Robust Estimation of Mahalanobis Distances in Hyperspectral Images

DISSERTATION

Eduardo C. Meidunas, Maj, USAF

AFIT/DS/ENG/07-02

DEPARTMENT OF THE AIR FORCE
AIR UNIVERSITY

AIR FORCE INSTITUTE OF TECHNOLOGY

Wright-Patterson Air Force Base, Ohio

The views expressed in this dissertation are those of the author and do not reflect the official policy or position of the Department of Defense or the United States Government.

AFIT/DS/ENG/07-02

Robust Estimation of Mahalanobis Distances in Hyperspectral Images

DISSERTATION

Presented to the Faculty

Graduate School of Engineering and Management

Air Force Institute of Technology

Air University

Air Education and Training Command

In Partial Fulfillment of the Requirements for the

Degree of Doctor of Philosophy

Eduardo C. Meidunas, B.S. Physics, M.S.E.E.

Maj, USAF

December 2006

AFIT/DS/ENG/07-02

Robust Estimation of Mahalanobis Distances in Hyperspectral Images

Eduardo C. Meidunas, B.S. Physics, M.S.E.E.

Maj, USAF

Approved:

Date

Dr. Steven C. Gustafson Dissertation Advisor	
Dr. Jeffrey P. Kharoufeh, Dean's Representative	
LtCol Matthew E. Goda, PhD Committee Member	
LtCol Robert E. Neher, PhD Committee Member	
Dr. Dimitris G. Manolakis Committee Member	

Accepted:

M. U. Thomas Dean, Graduate School of Engineering and Management	Date

Abstract

Hyperspectral images typically have a large number of band components, over 200 bands for many sensors, and thus the image pixels can be processed as data points in a multidimensional space. The processed outputs are generally represented as classification maps, or target detection/material identification maps and/or material abundance information maps. However, variable scene constituents and sensor limitations constrain the accuracy of the processed outputs. Specifically, stochastic hyperspectral processing techniques are limited by inaccurate models of the data. Previous work has shown that Mahalanobis distance distributions of the data govern the performance of stochastic models of the image pixels. Robust models of the tails of these distributions enable accurate threshold selections for detection and classification algorithms.

This dissertation develops new estimation methods that fit Johnson distributions and generalized Pareto distributions to hyperspectral Mahalanobis distances. The Johnson distribution fit is optimized using a new method which monitors the second derivative behavior of exceedance probability to mitigate potential outlier effects. This univariate distribution is then used to derive an elliptically contoured multivariate density model for the pixel data. The generalized Pareto distribution models are optimized using a new two-pass method that estimates the tail-index parameter. This method minimizes the mean squared fitting error by correcting parameter values using data distance information from an initial pass. A unique method for estimating the posterior density of the tail-index parameter for generalized Pareto models is also developed. Both the Johnson and Pareto distribution models are shown to reduce fitting error and to increase computational efficiency compared to the standard model of a mixture of F-distributions.

Acknowledgements

First and foremost, I wish to express my sincerest gratitude and appreciation toward my wife and my children. For my wife, your confidence in me and your understanding and patience with what I had to do to get this done has been the key to my success. I owe much of my professional accomplishment to you. For my children, I thank them for their patience and love and their ability to make me smile even during the toughest portions of this endeavor.

I would like to express my appreciation to Dr Gustafson for his tireless efforts as my research advisor. Thank you for asking the difficult questions and keeping me on my toes throughout this process. I would also like to thank Dr Manolakis for his vision and keeping me on the right track to compete this research. Thank you for presenting me with the ideas behind approaching this problem, and I greatly appreciate all of the time and effort you spent with me during the many teleconference meetings we had to shape this research. Thank you LtCol Goda for your insight and guidance in developing the final stages of this research. Thank you LtCol Neher for your mathematical expertise and for your conviction in the validity of this work. To each of my committee members: it has been my honor to have you as part of my Ph.D. committee. I hope to have the opportunity to work with you again in the future.

Finally, I would like to express my love and gratitude to my parents, who taught me the value of persistence and inspire me in all of my endeavors. I dedicate this work to you. To family members and friends who have shared in my plight, I thank you for your love and support. Without you all, this journey would not be as rich.

Eduardo C. Meidunas

Table of Contents

Page

List of Figures

Figure Page

List of Tables

List of Symbols

Symbol		Page
$\mathbf{x}$	HSI d-dimensional pixel	9
$\mathbf{e}$	d-dimensional endmember (pure material pixel spectrum)	9
a_k	abundance coefficient	9
$\mathbf{n}$	d-dimensional noise vector	9
$f_d(\mathbf{x} \mid \Theta_k)$	d-dimensional unimodal density for class k	14
Θ_k	parameter vector for density of material class k	14
$f_d(\mathbf{x})$	d-dimensional finite mixture model for pixel $\mathbf{x}$	14
$\boldsymbol{\mu}$	mean vector $\equiv E(\mathbf{x})$	14
$\boldsymbol{\Gamma}$	covariance matrix $\equiv E((\mathbf{x}-\boldsymbol{\mu_k})^T(\mathbf{x}-\boldsymbol{\mu_k}))$	14
$\mathbf{z}$	noise-whitened HSI data pixel (vector)	19
Δ^2	Mahalanobis distance	30
S_U	Unbounded Johnson distribution	46
z	standard normal scalar variate	46
ϵ_J	Johnson system location variable	46
γ_J	Johnson system shift variable	46
η_J	Johnson system shape variable	46
λ_J	Johnson system scale variable	46
S_L	single-side bounded Johnson distribution	46
S_B	Two-side bounded Johnson distribution	46
$S_d(\phi)$	d-dimensional spherically symmetric distribution	71
ϕ	characteristic generator	71
$\psi(t)$	characteristic function of a spherical distribution	72
Ω	$d \times d$ orthonormal operator	72
$N_d(\mathbf{0}, \mathbf{I}_d)$	d-dimensional normal distribution	72
γ_*	$\gamma_J - \eta_J \; ln(\lambda_J)$	75

Symbol		Page
μ	GEV (GPD) location parameter	83
σ	GEV (GPD) scale parameter	83
k	GEV (GPD) tail-index parameter	83
u	Threshold for extreme value sub-sets (defines how many of the largest values of an ordered set are to be used)	84

List of Abbreviations

Abbreviation		Page
HSI	Hyperspectral imaging	1
MD	Mahalanobis distance	4
GLRT	Generalized Likelihood Ratio Test	16
CFAR	Constant False Alarm Rate	21
ROI	Region of Interest	21
MLE	maximum likelihood estimate	27
ROC	Receiver Operating Characteristic	30
EC	Elliptically Contoured	33
SEM	stochastic expectation maximization	34
G	Cumulative Distribution Function	38
MSE	Mean Squared Error	39
GPD	generalized Pareto distribution	44
LOOS	Leave-One-Out-Smoothness	62
MCD	Minimum Covariance Determinant	62
PDF	probability density function	70
EVT	Extreme Value Theory	82
GEV	Generalized Extreme Value	83
RVs	random variables	91
POT	Peaks Over Threshold	92
MAP	maximum a posteriori	94
MMSE	minimum mean square error	94
MOM	Method of Moments	110
PWM	Probability Weighted Moments	112
EPM	Elemental Percentile Method	113
RMSE	Root Mean Squared Error	117

Abbreviation		Page
AMSE	Asymptotic Mean Squared Error	122
ROIs	Region of Interest	154
SPFC	South Panels Field Cropped	156
ENVI	Environment for Visualizing Images	156
MFC	Mixed Forest Cropped	156
MCFC	Mixed Coniferous Forest Cropped	157
DFC	Deciduous Forest Cropped	157
CFC	Coniferous Forest Cropped	157
ALPPC	All Loblolly Pine Plantations Cropped	157
CFCR	Coniferous Forest Cropped Reduced	157
ALPPCR	All Loblolly Pine Plantations Cropped Reduced	157

Robust Estimation of Mahalanobis Distances in Hyperspectral Images

I. Introduction

Remote sensing of earth objects from airborne platforms and space is of particular interest to the U.S. Air Force and the Department of Defense (DoD). Specifically, Hyperspectral imaging (HSI) is of great interest to the DoD due to its inherent ability to discriminate materials across a wide range of imaging wavelengths and its accessability to sub-pixel resolution processing. This research investigates the statistical characterization of HSI pixel data in order to develop the accurate stochastic models necessary for robust target/anomaly detection algorithms and for other statistical HSI image processing applications.

1.1 Hyperspectral Remote Sensing

Fundamentally, HSI is imaging spectroscopy from a distance [53]. Reflected solar irradiance data is gathered at discrete wavelengths in order to construct a spectrum for each pixel. Each material on the ground exhibits a unique spectrum based on how the material reflects and absorbs sunlight (see Figure 1.1). The resulting pixel spectrum demonstrates characteristics of the combined spectra of each inclusive material.

HSI offers high spectral resolution over a broad range of wavelengths, commonly from 0.4 to 2.5 microns. Typically, the spectral resolution samples every 9 to 12 nanometers, depending on the sensor, giving HSI sensors an excellent capability to not only identify different material classes but also to distinguish between materials in each class. This capability is primarily due to the fact that individual materials which occur within a class express variations in composition as slight shifts in peaks and troughs within a spectral curve continuum. As shown in Figure 1.1, a visual example of the fundamental HSI concept, different materials have distinct reflectance spectra.

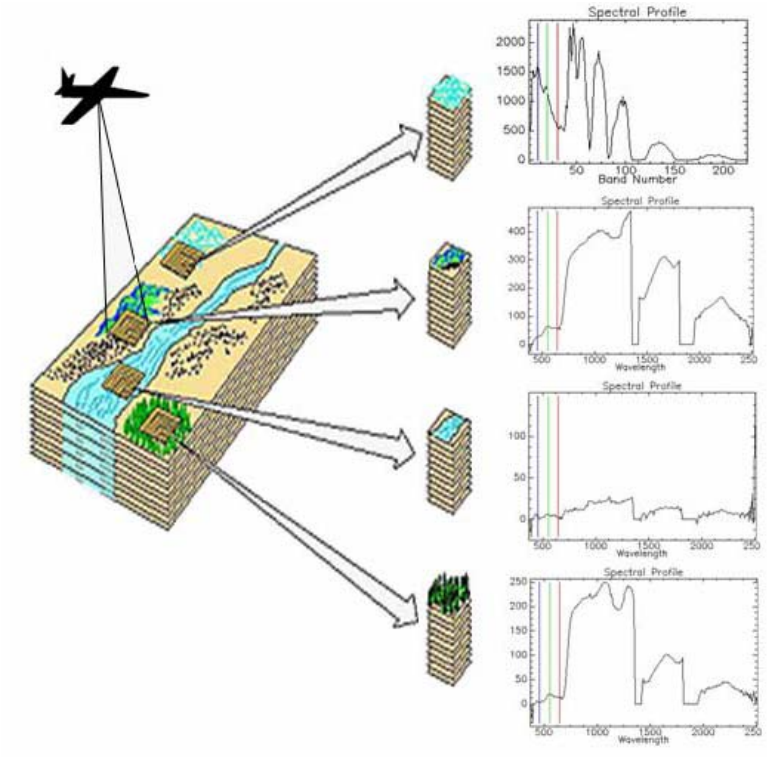


Figure 1.1: A diagram of the basic concepts in hyperspectral imaging. An image is sampled over a few hundred discrete wavelengths so that each pixel represents a radiance spectrum of its constituent material [24].

HSI data is collected in discrete data sets called hyper-cubes. A series of spatially continuous hyper-cubes constitutes a single run. Normally, a few runs are made over a desired scene for multiple looks by the sensor. A typical HSI spectrometer collects on the order of ~ 512 lines by ~ 500 samples by ~ 200 wavelength bands per hyper-cube. Figure 1.2 depicts the geometry of a single hyper-cube.

Compared to HSI, multispectral sensing only samples at a few wavelengths, generally from 5 to 10 bands. The sampling resolution is much lower for multispectral imaging, on the order of 100 to 200 nanometers, and hence multispectral sensing is not conducive to material discrimination. A multispectral sensor is not a true imaging spectrometer, as it does not offer a continuous spectral representation of the target being imaged, only a mapping of pixel data at discrete bands, where the high number and higher resolution of HSI sampling bands results in a near-continuous spectrum.

Thus multispectral data sets do not offer the processing opportunities of HSI sensors. The work considered here deals exclusively with HSI data exploitation.

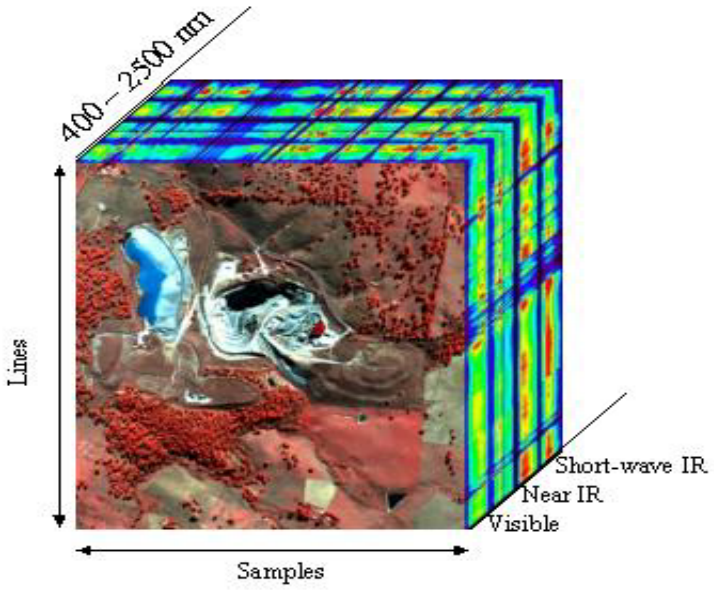


Figure 1.2: Geometry of a typical hyperspectral image cube.

1.2 Hyperspectral Imaging Systems

Hyperspectral imagery from the AVIRIS (Airborne Visible/Infrared Imaging Spectrometer) sensor is used in this work. AVIRIS is a state-of-the-art airborne HSI sensor developed by NASA Jet Propulsion Laboratory with a spectral coverage of 0.4 to 2.5 μm. The instrument samples over this wavelength range using 224 bands, each at a spectral resolution of 10 nm per band. A typical AVIRIS hyper-cube contains 512 x 512 pixels, approximately covering a 10 x 10 km swath, at a ground instantaneous spatial resolution of 20 x 20 m for high altitude imaging [24]. At lower altitudes the spatial resolution can be 1 x 1 m for a swath size of roughly 0.5 x 0.5 km.

The imagery used here is from a 1999 AVIRIS scene over Fort AP Hill, Virginia, which is an area that has been studied extensively using different HSI sensors. The site is a target-rich layout which is heavily "ground truthed" under various environmental conditions and sensor configurations, making it an attractive data set for HSI research.

Future work proposed here will use more recent AVIRIS data collects on this site as well as data sets by similar sensors over the same site.

1.3 Hyperspectral Image Processing

HSI data allows for great versatility in processing and exploitation. Since the pixels each have a large number of band components (over 200 bands for AVIRIS), information can be modeled and processed as data points in a multidimensional vector space. Signal processing techniques that combine matrix manipulation and stochastic modeling are the most common approaches for exploiting HSI data. Common HSI applications include (but are not limited to) target/anomaly detection and scene classification for military purposes, precision agriculture, global change detection, geology/mining, and space exploration.

There are many algorithms that achieve useful HSI data exploitation [6,30,41,53, 73]. These routines involve generally either geometric data processing or stochastic data processing methods. The research described here focuses on models that are involved in methods for stochastic processing.

1.4 Problem Statement

Stochastic data exploitation algorithms require accurate models of the underlying data statistics. Recent research shows that heavy-tailed distributions, i.e., probability distributions for which the corresponding probability density function decays slowly (typically following a polynomial instead of an exponential decay law), model HSI scene information well. Most importantly, it has been shown that the performance of detection algorithms is directly tied to accurately modeling the distribution of Mahalanobis distances (MD) of data comprising the background scene information in HSI [49].

MDs are contours of equal distance from the pixel points in a high-dimensional image cube with non-unit covariance to the centroid of the points. In contrast, Eu-

clidean distance is distance to the centroid of a cube with unit covariance. Euclidean distance becomes MD after the axes are linearly transformed by a non-unit covariance matrix, as shown in Figure 1.3. The MDs of HSI background data are of interest because, as explained in Chapter II, MD values above a threshold are important for modeling detection performance.

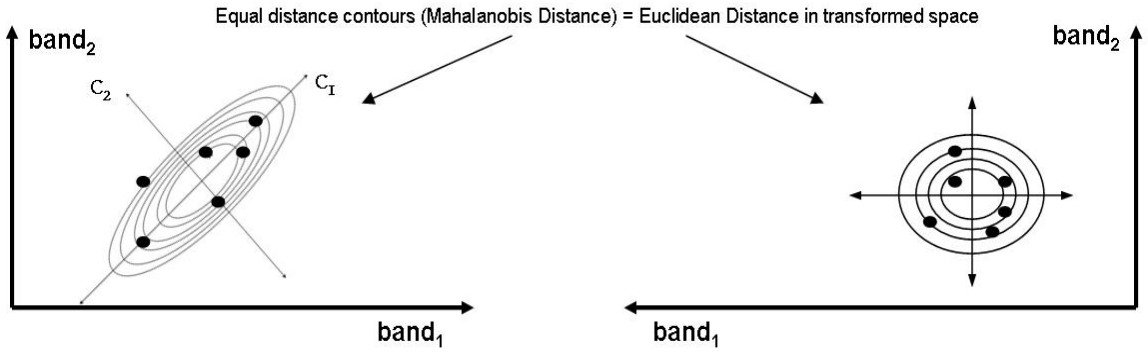


Figure 1.3: The MD as a measure determined by the covariance of the data points.

Proper modeling of the MD distribution allows for accurate detector threshold selection at a given false alarm rate, as shown in Figure 1.4. However, scene variability introduces substantial error in estimating the parameters that describe heavy-tailed MD distributions. Thus robust methods for estimating the parameters which describe heavy-tailed distribution models in HSI are needed.

A hypothesis of this research is that heavy-tailed distribution models of HSI data, particularly MD distributions from clusters of data, can be improved and optimally constructed with parameters robustly estimated using information from the data. Each of the objectives associated with this hypothesis are listed below and detailed in the next chapter. The heavy-tailed distribution models used in this research are Johnson distributions and generalized Pareto distributions (GPDs).

Objective 1: Determine a robust Johnson distribution model for HSI MD data: Determine a Johnson distribution model for heavy-tailed HSI MD distributions that is robust to possible outlier and/or anomaly perturbations.

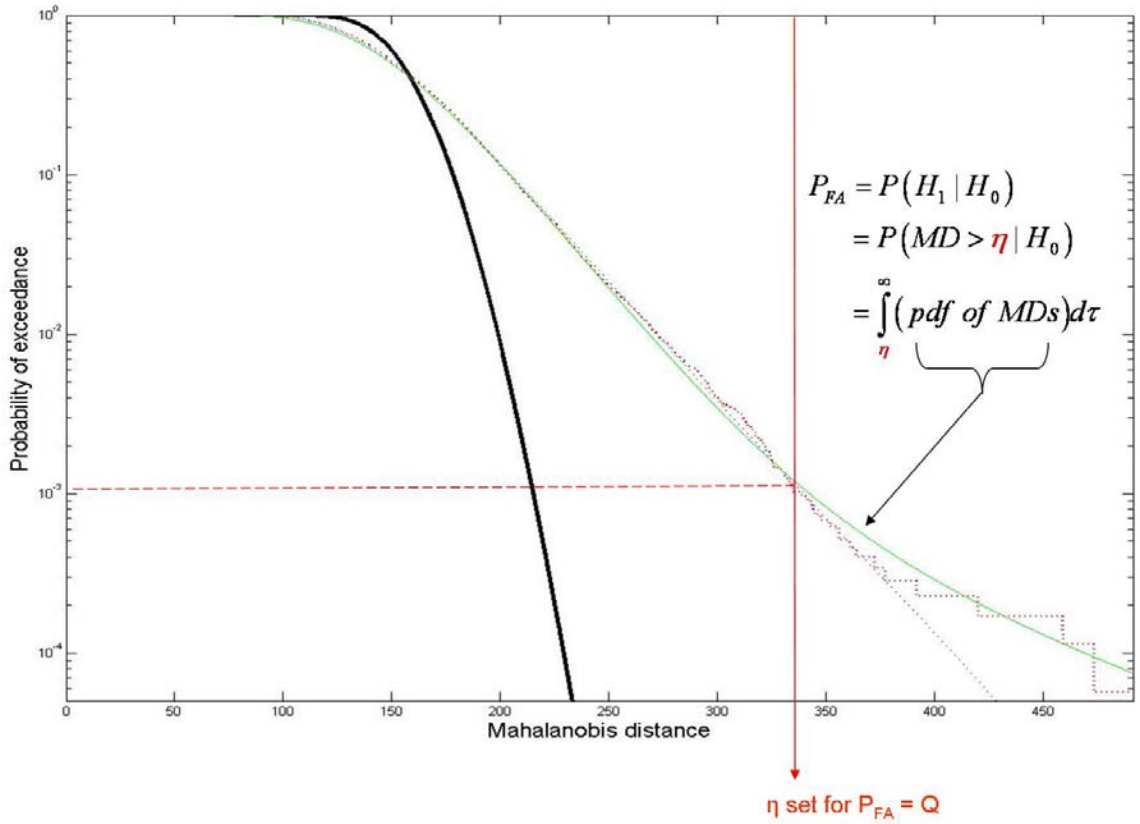


Figure 1.4: The probability density function that properly models background (H_0) as opposed to target (H_1) HSI data optimizes a detection routine by allowing accurate threshold (η) selection for a specified false alarm probability ($P_{FA} = Q$). Equating $P(H_1|H_0)$ to $P(MD > \eta|H_0)$ is described in Chapter II. Here the MD data is a dashed line and three models for the data are given by the dark solid line, light solid line, and dotted line; these models are explained in Chapter III. Probability of exceedance at a given MD is $1 - CDF$, where CDF represents the Cumulative Distribution Function of the MDs. For this example, the probability of exceeding MD = 340 is 0.001. Therefore, for a desired $P_{FA} = 0.001$, a cutoff is set for identifying pixels with MDs above 340.

Objective 2: Determine a multivariate elliptically contoured density model from the univariate Johnson distribution: Once the univariate Johnson distribution for describing HSI MD distributions is found, apply multivariate symmetric distribution theory to derive the multivariate elliptically contoured HSI data density.

Objective 3: Determine a viable parameter estimation method for obtaining tail-index values for GPD models of MD: Analyze and assess a suite of tail-index estima-

tion methods for GPDs that model MD distributions. Perform a sensitivity analysis of each model with respect to changing MD data size. Identify the best method for estimating tail-index parameters for GPDs applied to heavy-tailed HSI MD data.

Objective 4: Develop an optimization method for obtaining robust tail-index estimates for GPD models: Apply a feedback mechanism to the best tail-index parameter estimation method identified in Objective 3. Based on the feedback, optimize, with respect to the affect of possible outliers/anomalies, the estimator for a resulting GPD fit to MD distribution data to obtain decreased mean-squared error.

Objective 5: Assess the utility of Johnson and GPD models for stochastic HSI processing: Assess the relative utility of using Johnson distribution models for heavy-tailed HSI MD behavior and optimized Pareto models.

1.5 Dissertation Outline

The development and rationale for the objectives is covered in Chapter II. The methodology roadmap and overview for the current research is also discussed in Chapter II. Chapters III-VI consider the specifics of each objective. Finally, Chapter VII reviews the objectives, discusses recommendations, and highlights the contributions of this research.

II. Background on HSI Stochastic Processing Methods

This chapter provides the background on hyperspectral data modeling and methods for hyperspectral data exploitation. It introduces the framework for stochastic HSI processing and explains the significance of fitting MD distributions accurately. It then describes the most recent model for MD behavior.

2.1 Hyperspectral Data Model

Hyperspectral imaging sensors passively measure the radiance of materials within the field of view of the optics, with each pixel area sampling at a large number of contiguous bands. This operation is summarized in Figure 2.1, which shows different scene constituents observed by an HSI sensor, and Figure 2.2, which relates the scene sampling process to an output of a spectrum of contiguous bands.

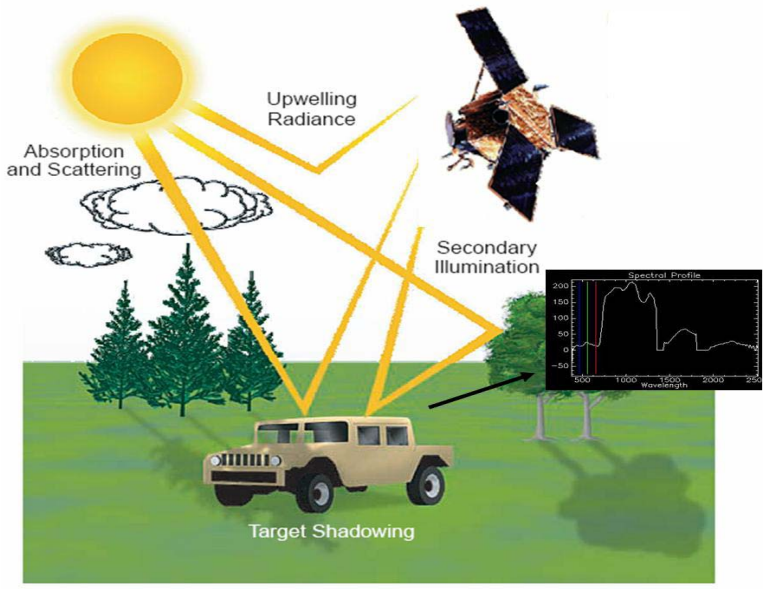


Figure 2.1: Hyperspectral imaging architecture and pixel spectrum example (from [55] and modified to show a resulting pixel spectrum).

The contiguous bands may be viewed as dimensions, and the resulting pixel spectra may be viewed as a vector in the multidimensional space. The architecture of the hyperspectral image is then interpreted using a linear mixing model

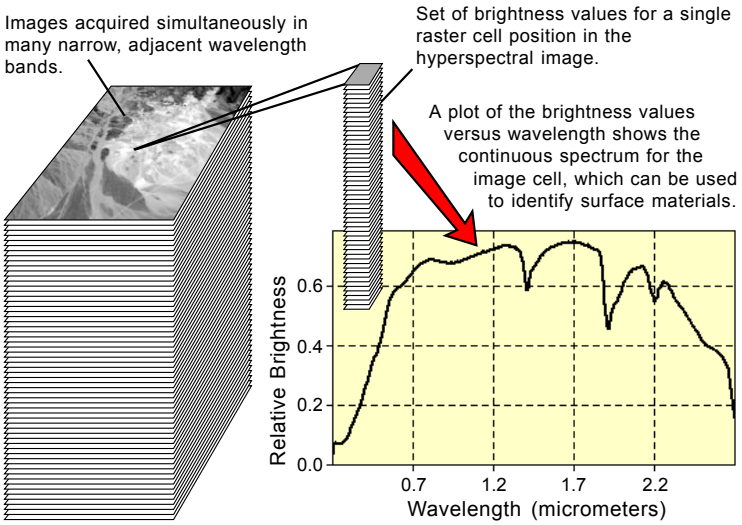


Figure 2.2: Relationship of pixel components to reflectance profile (from [63]).

$$\mathbf{x} = \sum_{k=1}^{N} a_k \mathbf{e}_k + \mathbf{n}, \tag{2.1}$$

where $\mathbf{x}$ is the pixel spectrum (d-dimensional data vector), $\mathbf{e}$ is the k^{th} endmember (the N endmembers are basis vectors that linearly combine to define any data point in the d-dimensional space), a_k is the mixing coefficient (or abundance value) corresponding to the k^{th} endmember, and $\mathbf{n}$ is a d-dimensional noise vector due to atmospheric scattering and absorption, secondary illumination, shadowing, pixel resolution, sensor configuration, and scene variability. The matrix-vector form,

$$\mathbf{x} = \mathbf{E}\mathbf{a} + \mathbf{n}, \tag{2.2}$$

may be solved for $\mathbf{a}$ using matrix inversion techniques.

2.1.1 Geometric Model. Paralleling the linear mixing model is a geometric interpretation of the above equations. Boardman *et al.* explain the d-dimensional space of the HSI data with a convex hull analogy [6]. In this model the pixel data are data points residing within a d-dimensional convex hull inscribing the data space. The

endmembers are the vertex data points defining the shape of the multidimensional simplex (hence the term "endmembers" - they literally reside at the "ends" of the space inscribed by the convex hull). This model is illustrated in Figure 2.3.

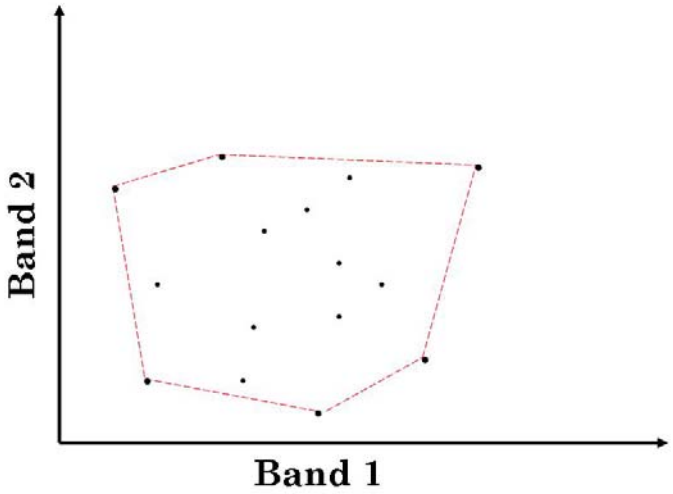


Figure 2.3: A simple example of a convex hull with endmembers. A simplified data space resulting from an HSI system using only two bands is shown. The convex hull is depicted by the dashed line and the endmembers are the data points at the vertices of the hull.

This geometric approach facilitates orthogonal subspace projection and oblique subspace projection techniques. In both cases the HSI data is projected onto a subspace of reduced dimensionality in which the pixels are identified by the position of their vectors in the projection space [27]. Subspace projection algorithms are typically used for target/anomaly detection because anomalously shaped pixel spectra "stand out" when projected against a subspace where the majority other pixels cluster around a specific locus. An example of subspace projection in the geometric context is shown in Figure 2.4.

Geometric models perform well under near full-pixel mixing scenarios, i.e., for image pixels which contain a majority of one type of material spectra. The resolution of the sensor is thus sufficient to enable integration of the spectra from a scene that results in the majority of pixels in the image containing near "pure" spectra, where

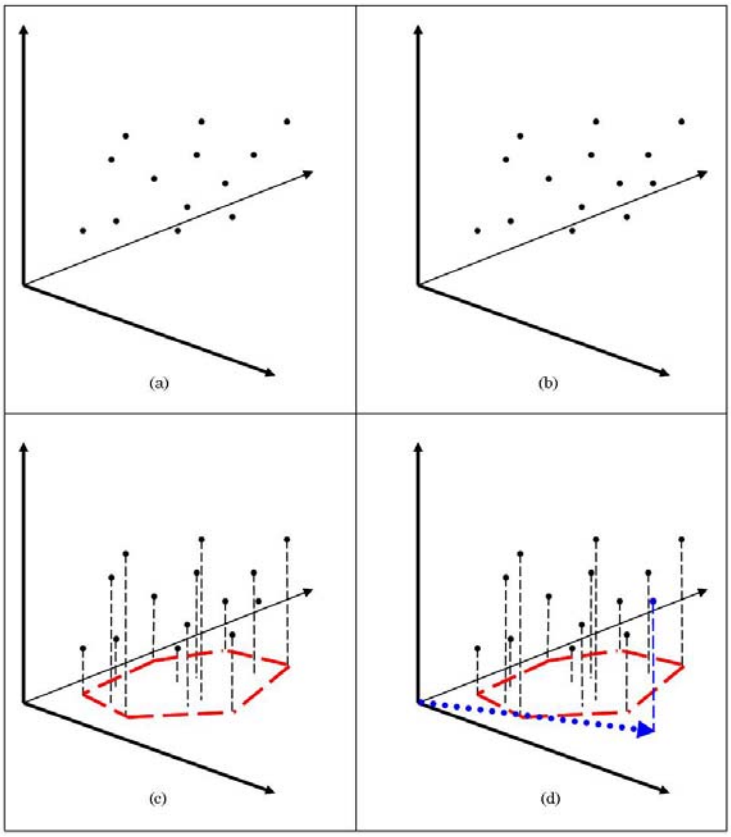


Figure 2.4: An example of subspace projection in terms of geometric manipulation. In (a) an original data set is represented in three dimensions (similar to viewing HSI data using only three bands). In (b) a data point is added to the original data set. Notice that it is difficult to discern whether the added point belongs to the same family as the original data set. In (c) the original data are projected onto a two dimensional subspace and a convex hull is drawn to delineate the boundaries of the data set. In (d) it is clear that the additional point projection is outside of the boundaries specified for the original data set and is therefore anomalous to the original data set in the projection space.

a "pure" pixel spectrum is an all grass spectrum, an all asphalt spectrum, etc. With high spatial resolution sensors, the majority of image pixels contain large abundances of only a few materials (if not containing 100 percent of the spectrum of a single material), thus being near-pure. Under such scenarios, the endmembers extracted from a scene are independent (or nearly independent), and the matrix inversion and subspace projection techniques are effective.

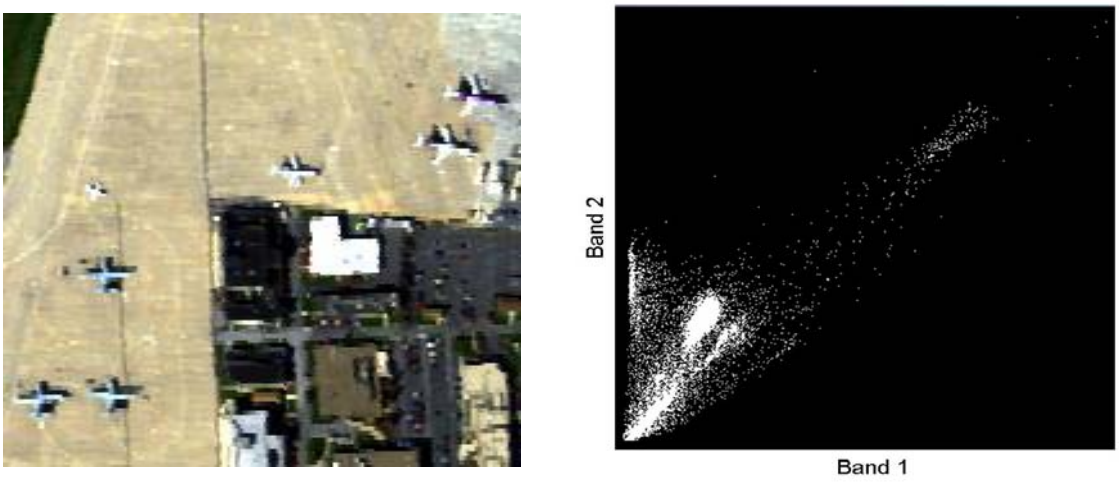
Geometric models become ineffective once a high degree of sub-pixel mixing occurs. In such scenarios the sensor resolution is low and the integration of many image spectra are realized in one pixel. The sub-pixel mixing of spectra is random, and greater variability within the scene creates greater pixel-to-pixel mixing variability. Also, since the majority of the pixels are mixtures of spectra from many materials taken from the scene, the endmembers are highly correlated (as are the pixels in the scene) and, therefore, not independent. Matrix inversion and projection techniques are ineffective under these conditions, and geometric models, and the techniques which manipulate the geometry of the scene cannot characterize the wide range of pixel variability.

2.1.2 Stochastic Model. Statistical variability in a scene is better described using a probabilistic model. The range of d-dimensional pixel values from a d-dimensional HSI image can be viewed as a scatter plot of points in a d-dimensional space. The points in this space are the endpoints of vectors defined by the d-dimensional component values at every HSI wavelength band.

Given a HSI image with only two wavelength bands (the "wavelength" term will henceforth be eliminated and "bands" will refer to "wavelength bands"), the endpoints of the pixel vectors may be modeled as a two-dimensional scatter plot. Figure 2.5 (b) shows the two-dimensional data from 40,000 pixels taken from the HSI image in Figure 2.5 (a).

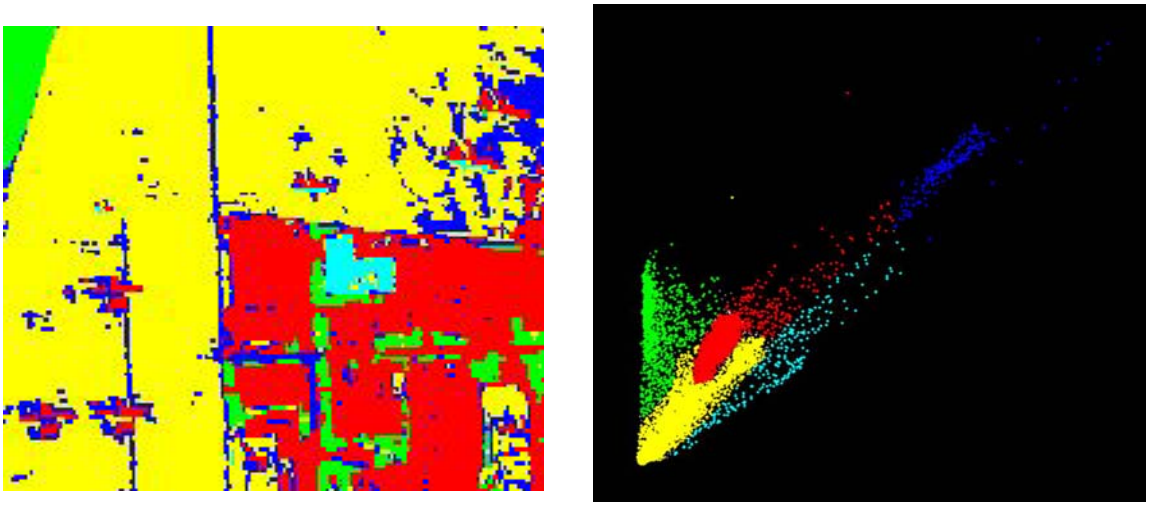
Areas where pixels have similar data characteristics are seen to cluster into groups in the scatter plots. Figure 2.6 (a) shows the image classified using simple K-means clustering, and Figure 2.6 (b) shows the actual data clustered in the two-dimensional band space. In K-means clustering the points are assigned to candidate cluster means, the means are re-calculated, and iteration continues until the cluster means do not change [19].

In the statistical model, clusters of similar data are assumed to be homogeneous, d-dimensional, multivariate distributions. Also, a specific material originates



(a) (b)

Figure 2.5: (a) AVIRIS HSI image containing 40,000 pixels of spectral data from the Harrisburg, PA airport [24]. (b) Scatter plot of the data values from (a) for two bands. The pixel values appear as points at the two-dimensional band coordinates.



(a) (b)

Figure 2.6: (a) Classification of image in Figure 2.5 (a) using K-means clustering. (b) Scatter plot demonstrating clustering of similar data. Each color represents a unique class.

from a unimodal d-dimensional multivariate distribution, which describes all possible variations of the material. A HSI image is a mixture of these material distributions, resulting in a multi-modal d-dimensional multivariate distribution. Each pixel in the image is a random vector from the multi-modal parent distribution associated with the image. A single pixel is assumed to originate from a mixture probability density function

$$f_d(\mathbf{x}) = \sum_{k=1}^{N} a_k f_d(\mathbf{x} \mid \Theta_k), \tag{2.3}$$

where $f_d(\mathbf{x} \mid \Theta_k)$ is the d-dimensional, uni-modal component distribution, with parameters Θ_k from homogeneous material class k and $f_d(\mathbf{x})$ is a mixture of the k classes with probability a_k [54].

Usually the finite mixture model of Equation (2.3) is constructed using multivariate normal distributions for the k component distributions. In this case $f_d(\mathbf{x} \mid \Theta_k)$ is

$$f_d(\mathbf{x} \mid \Theta_k) = \left((2\pi)^{d/2} |\boldsymbol{\Gamma_k}|^{1/2}\right)^{-1} \exp\left(-\frac{1}{2}(\mathbf{x} - \boldsymbol{\mu_k})^T \boldsymbol{\Gamma_k}^{-1} (\mathbf{x} - \boldsymbol{\mu_k})\right), \tag{2.4}$$

with the parameters Θ_k consisting of mean vector $\boldsymbol{\mu} \equiv E(\mathbf{x})$ and covariance matrix $\boldsymbol{\Gamma}$ $\equiv E((\mathbf{x} - \boldsymbol{\mu})^T (\mathbf{x} - \boldsymbol{\mu}))$. The quadratic expression in the exponent of Equation (2.4) is the squared Mahalanobis distance [19]

$$\Delta^2 = (\mathbf{x} - \boldsymbol{\mu_k})^T \boldsymbol{\Gamma_k}^{-1} (\mathbf{x} - \boldsymbol{\mu_k}). \tag{2.5}$$

For a multivariate (d-dimensional) distribution, contours of constant density are hyperellipsoids of constant Mahalanobis distance. Figure 2.7 shows a three-dimensional distribution ellipsoid with its contours of constant probability projected onto one of the planes. These contours represent a projection of Mahalanobis distances.

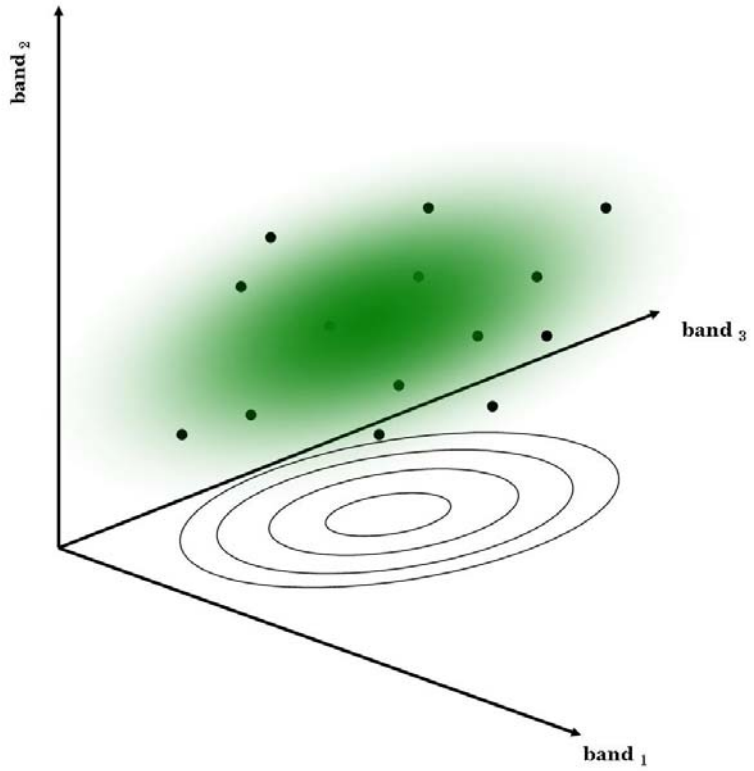


Figure 2.7: A set of data points in three dimensions. The data set is represented by a data cloud modeled by a three-dimensional Gaussian density. The Mahalanobis distance contours are projected onto the two-dimensional plane below it. These concentric ellipses represent contours of equal probability for the Gaussian hyperellipsoid.

This statistical distance measure is used in the majority of statistical detection algorithms.

2.2 Statistical Hyperspectral Signal Processing

Stochastic hyperspectral detection algorithms (whether detecting targets or identifying anomalies) are divided into two categories: full-pixel detectors and subpixel detectors. In the following, "target" describes any signal of interest not considered part of the natural background of the scene (whether from a man-made source or an anomalous source) . Full-pixel targets do not contain contamination resulting from the interaction of target and background spectra. Here "background" spectra refers to all non-target pixel spectra (usually from naturally occurring scene constituents).

2.2.1 Full-pixel Signal Processing. The statistical detection problem is formulated as a classical binary hypothesis test: $\boldsymbol{H_0} \Rightarrow$ *Target not present (background only)*, $\boldsymbol{H_1} \Rightarrow$ *Target present (target and background).* In HSI the two hypotheses involve knowledge of unknown parameters from the data set (covariance matrix and mean vector), which requires that the parameters be estimated and that the target detector be adaptive. However, the size of the background data set compared to the target data set creates conditions which make the estimation of these parameters challenging [51].

All remote sensing HSI target detection scenarios of interest to the DoD entail an architecture whereby the background data set is the majority of the image, while the target data set consists of only a few data points. The sparsely populated target set in the image offers little information for estimating its stochastic model parameters. Conversely, the large background data set, along with the target pixels in the image, necessitates the incorporation of the entire image pixel set for modeling background statistics. These constraints on the information in HSI lead to certain cases of target detection models, each of which is designed in a generalized likelihood ratio test (GLRT) framework [49].

The simplest approach models the target and background data sets as multivariate normal distributions. The detection hypothesis under this model is

$$\begin{aligned} \boldsymbol{H_0} &: \mathbf{x} \sim N(\boldsymbol{\mu_0}, \boldsymbol{\Gamma_0})\ \textit{Target not present}\ (\textit{background only}) \\ \boldsymbol{H_1} &: \mathbf{x} \sim N(\boldsymbol{\mu_1}, \boldsymbol{\Gamma_1})\ \textit{Target present}\ (\textit{target and background}), \end{aligned}$$

where the zero subscript pertains to the background model and the unit subscript refers to the target present model. Assigning conditional probability density functions to $\mathbf{x}$ under each hypothesis, the likelihood ratio is

$$\Lambda(\mathbf{x}) = \frac{p\left(\mathbf{x} \mid H_1\right)}{p\left(\mathbf{x} \mid H_0\right)}. \tag{2.6}$$

Using the GLRT methodology, a threshold η is set such that if $\Lambda(\mathbf{x})$ exceeds η, H_0 is rejected. A Neyman-Pearson detector is constructed using the known stochastic information under each hypothesis. When the covariances are not equal, $\mathbf{\Gamma_0} \neq \mathbf{\Gamma_1}$, the likelihood ratio is

$$\Lambda(\mathbf{x}) = \frac{|\mathbf{\Gamma_1}|^{1/2} \exp\left(-\frac{1}{2}\left(\mathbf{x}-\boldsymbol{\mu_1}\right)^T \mathbf{\Gamma_1}^{-1}\left(\mathbf{x}-\boldsymbol{\mu_1}\right)\right)}{|\mathbf{\Gamma_0}|^{1/2} \exp\left(-\frac{1}{2}\left(\mathbf{x}-\boldsymbol{\mu_0}\right)^T \mathbf{\Gamma_0}^{-1}\left(\mathbf{x}-\boldsymbol{\mu_0}\right)\right)}, \tag{2.7}$$

and the decision statistic in terms of the log-likelihood ratio $\ell(\mathbf{x})$ is

$$\begin{aligned} y &= 2\ell(x) \ln\left(\frac{|\mathbf{\Gamma_1}|^{1/2}}{|\mathbf{\Gamma_0}|^{1/2}}\right), & (2.8)\\ &= \left(\mathbf{x}-\boldsymbol{\mu_0}\right)^T \mathbf{\Gamma_0}^{-1}\left(\mathbf{x}-\boldsymbol{\mu_0}\right) - \left(\mathbf{x}-\boldsymbol{\mu_1}\right)^T \mathbf{\Gamma_1}^{-1}\left(\mathbf{x}-\boldsymbol{\mu_1}\right). & (2.9) \end{aligned}$$

The statistic in Equation (2.8) compares the difference in Mahalanobis distance of the test pixel spectra from the background mean and target mean. The value for y that maximizes the probability of detection subject to a specified probability of false alarm is selected for a Neyman-Pearson test. This test is depicted (in two dimensions) in Figure 2.8

The test in Figure 2.8 may be compared to the subspace projection example shown in Figure 2.4. The geometric model defines a target or anomaly as the signal which exceeds some boundary defined by the geometric configuration of the data set (the convex hull is given as one type of boundary). A direct analogy to the stochastic model can be made, where the boundary is now a Mahalanobis distance cutoff (or exceeding some boundary MD contour). This analogy is shown in Figure 2.9.

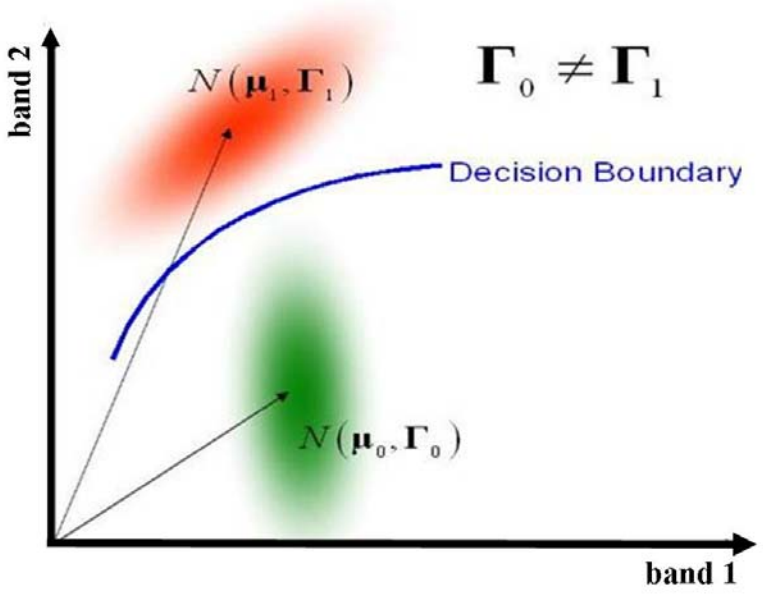


Figure 2.8: Target detector architecture for unequal covariances (the simplest case using the Generalized Likelihood Ratio Test for Neyman-Pearson criteria). The arrows to the centers of the data clouds are possible vector representations of points in the data cluster.

If the target and background classes have the same covariance matrix, then from Equation (2.8)

$$y = \vartheta \boldsymbol{\Gamma}^{-1} (\boldsymbol{\mu_1} - \boldsymbol{\mu_0}) \mathbf{x}. \tag{2.10}$$

The result is a Fisher Linear Discriminant [19], where ϑ is a normalizing constant. In most signal processing applications, Equation (2.10) is

$$y = \Phi_{MF}^T \mathbf{x}, \tag{2.11}$$

where the MF subscript indicates a matched filter operator [49]. The matched filter is illustrated in Figure 2.10.

The matched filter projects the test pixel spectra in the direction of Φ_{MF}^T. For optimal separability between the target and background clusters, Φ_{MF}^T must be in the direction of maximum distance between $\boldsymbol{\mu_0}$ and $\boldsymbol{\mu_1}$. The common signal processing

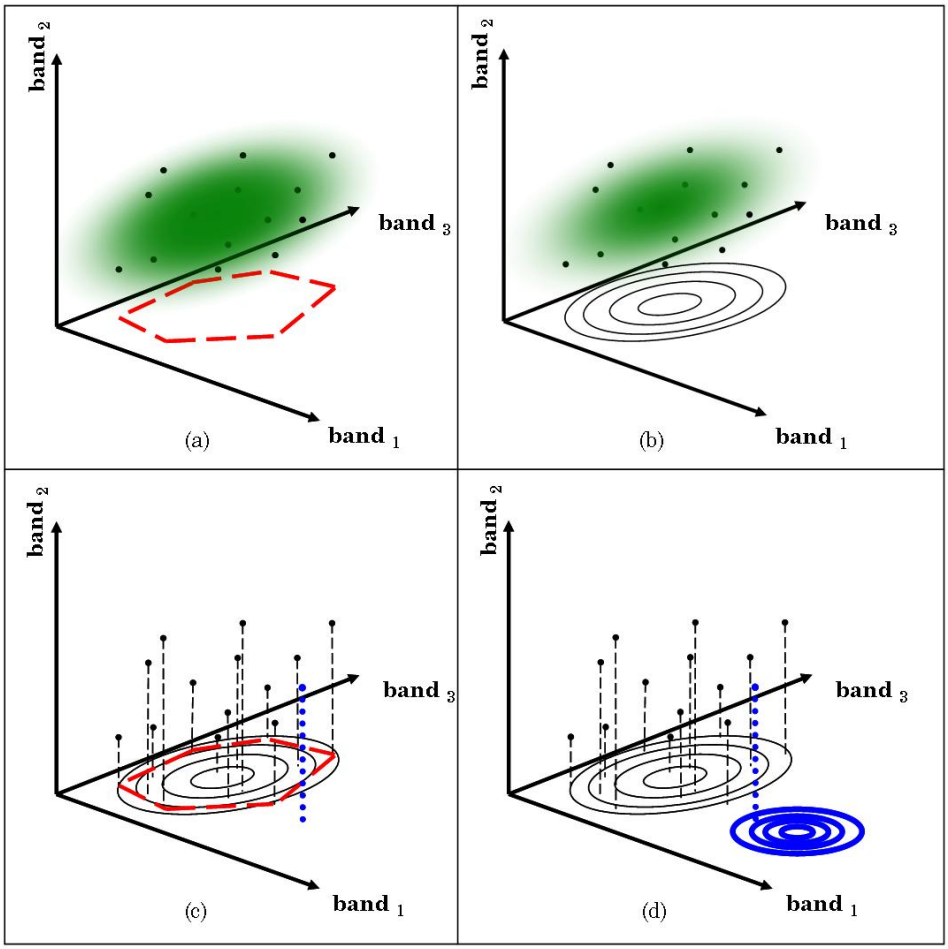


Figure 2.9: An analogy between the stochastic model and geometric model for HSI processing and the relationship between the Mahalanobis distance decision boundary in the statistical space as depicted in Figure 2.8 and the convex hull boundary in the geometric space shown in Figure 2.4. In (a) the data set is shown with a hyperellipsoid (representing a multidimentional Gaussian density model) and the convex hull projection. In (b) the hull is replaced with MD contours. In (c) the MD contours and convex hull are overlayed to demonstrate similarity and to compare distance to a potential anomalous data point. In (d) a representation of the anomalous data point belonging to another density is represented by different MD contours. The decision boundary shown in Figure 2.8 is located between the two MD contour groups.

practice of "whitening" performed on the data [38] achieves this effect, where the whitening parameter $\mathbf{\Gamma}^{1/2}$ is applied to $\mathbf{x}$ ($(\mathbf{z}) = \mathbf{\Gamma}^{1/2}\mathbf{x}$), the test pixel spectra, giving the detector architecture shown in Figure 2.11.

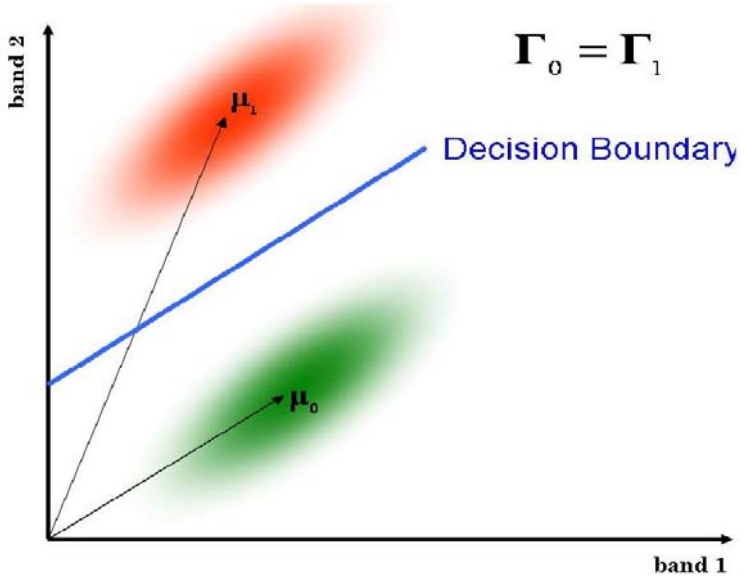


Figure 2.10: Target detector architecture for equal covariances, which is also the case that leads to matched filter detection.

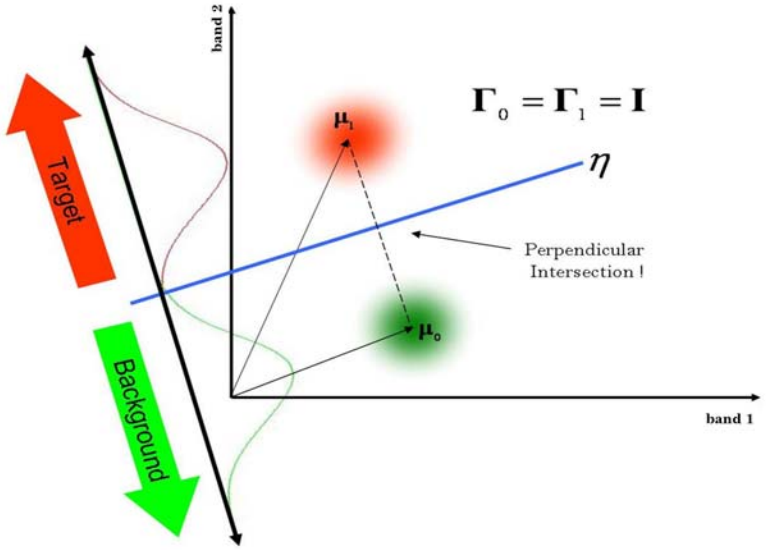


Figure 2.11: Target detector architecture for equal covariances and its transformation into whitened space, illustrating the optimal matched filter target detector concept.

As can be seen in Figure 2.11, the optimal matched filter detector measures the MD between the target and background data sets projected onto the dotted line perpendicular to the direction of maximum separation between the two sets. Under the

assumption of normality for the data, this is the optimal detector architecture. Such a detector exhibits constant false alarm rate (CFAR) performance, where CFAR means that the detector probability of false alarm is independent of the noise covariance matrix [40, 49].

However, the matched filter presented here requires sufficient a *priori* information about both data sets for a complete stochastic model. In many common HSI scenarios, not enough information is available to model the statistics of the target data set, and the only information available is the background mean and covariance. Using the GLRT approach and assuming normally distributed data for the simplest case, the target and background covariances are equal and the adaptive anomaly detector is constructed [78].

For the anomaly detector, the detection statistic, since only $\boldsymbol{\mu_0}$ and $\boldsymbol{\Gamma_0}$ are known, becomes the MD of the test pixel spectra from the background mean. The threshold, defined by the desired probability of false alarm, is a radial distance from the mean of the background data set in the image data space. An anomaly (possible target) is detected once the threshold is exceeded by the MD of the test pixel spectra from the background mean. The anomaly detector is shown in Figure 2.12.

For the majority of HSI, neither the background statistics nor the target statistics are known. In the target detection image scenario the target data sparsely populate the image, while the background data form the majority of the image data set. In this case the background statistics are estimated from the entire image data set. If only a certain region of interest (ROI) in the image is under consideration, the mean vector and covariance matrix of the background are estimated from the ROI, provided that the region is large enough to yield an invertible estimated covariance matrix and small enough to ensure homogeneity of the background. The target detector for this architecture is called the adaptive anomaly detector [40].

However, estimating background statistics in this manner results in the loss of Neyman-Pearson optimality in the detector, and selecting the proper ROI for es-

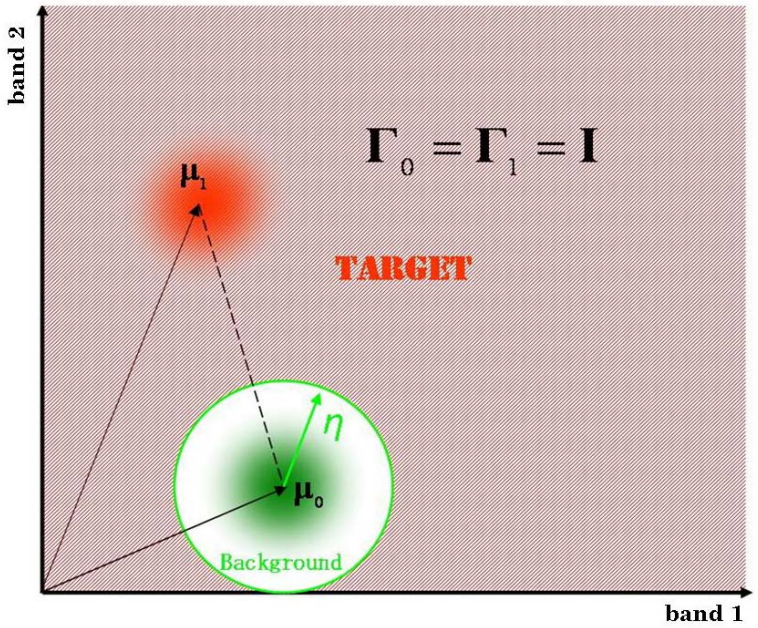


Figure 2.12: Anomaly detector for normally distributed data with equal covariances (transformed into whitened space). The line between the target data set and the background data set represents the Mahalanobis distance between the two sets. Any value exceeding the radial threshold around the background data set is classified as a target.

timating statistics becomes important. Optimality is approached as the quality of the estimated covariance matrix increases [49]. Also, if the covariance matrix is estimated using a data population size greater than three times the dimensionality of the HSI scene, and if normally distributed data is assumed, then the loss in detection performance is roughly 3 dB [40].

A clear conclusion, however, is that these detectors rely on a statistic governed by MD. Properly modeling the MD distribution allows for a proper model of the distribution of the detection statistic and hence accurate detector parameters (false alarm rates and detection probabilities). Note that the detectors developed in this sub-section depend on full-pixel signal fill. The next sub-section investigates the behavior of detectors under sub-pixel signal mixing.

2.2.2 Sub-pixel Signal Processing. Examples of target detectors in the previous section are greatly simplified due to the full-pixel target fill assumption. In practice, most hyperspectral imagery is taken from a high altitude and with a spatial resolution that results in target material spectra mixing with surrounding spectra in each pixel. This scenario, with the addition of noise from various sources such as the atmosphere, the sensor, and the scene composition, create multiple "sub-pixel" mixing components (individual spectra) for each pixel, representing the integration of different materials. Under these conditions, sub-pixel target detection becomes difficult.

In the stochastic model (sometimes referred to as the Bayesian model in the presence of noise [39]) for sub-pixel detection, the hypothesis test is altered to include these effects. The test is

$$\begin{aligned} \boldsymbol{H_0} &: \mathbf{x} = \mathbf{n} \quad \textit{Target not present } (\textit{background only}) \\ \boldsymbol{H_1} &: \mathbf{x} = \mathbf{E}\mathbf{a} + \mathbf{n} \quad \textit{Target present } (\textit{target and background}), \end{aligned}$$

where $\mathbf{E}$ is the matrix of endmembers or target data (in columns) representing the variability in the image data, $\mathbf{n}$ is the additive noise modeled as a normal distribution with zero mean vector and covariance matrix $\mathbf{\Gamma}$ [29, 40, 71], and $\mathbf{a}$ is the abundance vector (the vector of values corresponding to the amount of each endmember (columns in matrix $\mathbf{E}$) present in the pixel vector $\mathbf{x}$). This unstructured background model [49] is so-called due to background characterization as a single entity with environmental noise (atmospheric, sensor, illumination) not separated from the background mixing process. Using this model as shown in Figure 2.13, $\mathbf{x} \sim N(\mathbf{0}, \mathbf{\Gamma})$ under $\boldsymbol{H_0}$ and $\mathbf{x} \sim N(\boldsymbol{Ea}, \mathbf{\Gamma})$ under H_1.

In the unstructured background model it is assumed that the test pixels and endmembers are independent. Also, the spectra intermingle on the sub-pixel level

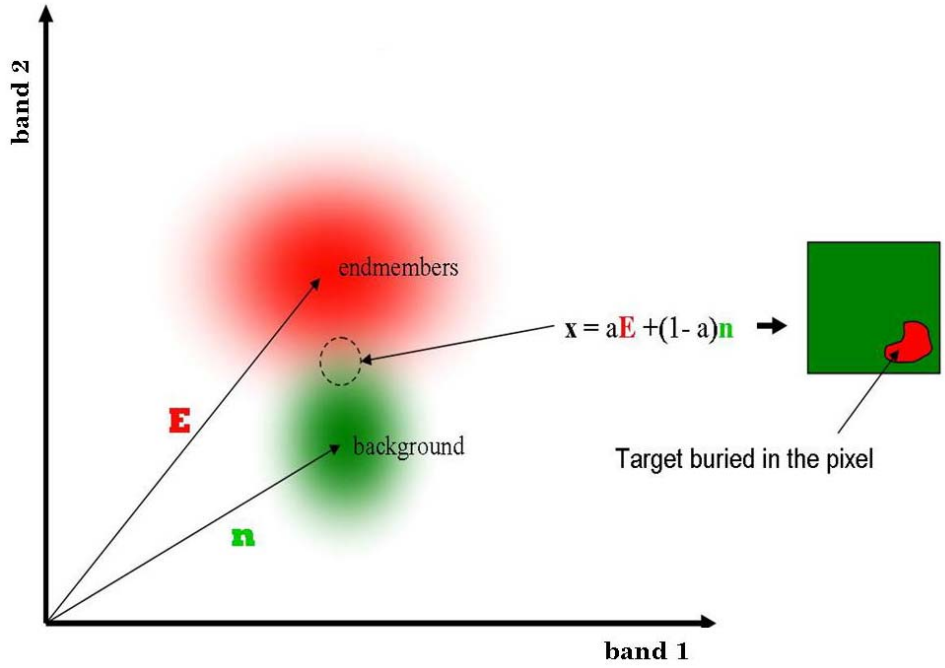


Figure 2.13: Sub-pixel mixture model. Pixels inside the dotted circle have varying proportions of background and endmember spectra (target spectra).

in an additive manner (i.e., endmembers and background pixel spectra are added in proportion and not replaced). Using the GLRT approach with this model and the maximum likelihood estimate of the covariance matrix $\hat{\mathbf{\Gamma}}$, the RX detector is derived with the statistic [65]

$$y = \mathbf{x}^T \left(\frac{N}{N+1} \hat{\mathbf{\Gamma}} + \frac{1}{N+1} \mathbf{x}\mathbf{x}^T \right)^{-1} \mathbf{x}, \tag{2.12}$$

where the RX detector is a CFAR adaptive anomaly detector [1] and is a common detection algorithm for HSI. As N becomes large, the RX detection statistic converges to

$$y = \mathbf{x}^T \hat{\mathbf{\Gamma}} \mathbf{x}, \tag{2.13}$$

which is the Mahalanobis distance from the zero-mean background data set.

The unstructured background model is still an oversimplification. The structure of spectral mixing within pixels is such that (since different pixels contain different abundances for each material present in the scene) the background term for each hypothesis is different. Under this modification, the hypotheses are

$$\begin{aligned} \boldsymbol{H_0} &: \mathbf{x} = \mathbf{n} \quad \textit{Target not present (background only)} \\ \boldsymbol{H_1} &: \mathbf{x} = \mathbf{Ea} + \sigma\mathbf{n} \quad \textit{Target present (target and modified background)}, \end{aligned}$$

where σ represents the differing background component present in the mixed pixel. This modification implies that the covariance matrix for the target and background are different. Using this model, $\mathbf{x} \sim N(\mathbf{0}, \mathbf{\Gamma})$ under $\boldsymbol{H_0}$ and $\mathbf{x} \sim N(\boldsymbol{Ea}, \sigma^2\mathbf{\Gamma})$ under H_1 [49].

Kraut, Scharf and McWhorter [43] address the above problem by constructing the CFAR Adaptive Subspace Detector for Coherent Detection. In the whitened space it is

$$y = \frac{\mathbf{z}^T \tilde{\boldsymbol{E}} \left(\tilde{\boldsymbol{E}}^T \tilde{\boldsymbol{E}} \right)^{-1} \tilde{\boldsymbol{E}}^T \mathbf{z}}{\mathbf{z}^T \mathbf{z}}, \tag{2.14}$$

where $\tilde{\boldsymbol{E}} = \hat{\mathbf{\Gamma}}^{-1/2}\mathbf{E}$. This detector is also called the Adaptive Cosine Estimator because the detection statistic compares a threshold to the cosine of the angle the test pixel makes with the target data in noise with an unknown covariance structure. It is one in a family of adaptive Matched Subspace Detectors they discuss [43], and, with the matched filter and anomaly detector, comprise three of the four adaptive detectors considered in their paper. The fourth adaptive detector deals with subspace detection in structured backgrounds.

Modifying the unstructured background model to fully define the effects of environmental noise and background mixing noise yields hypotheses

$$\boldsymbol{H_0} \quad : \quad \mathbf{x} = \mathbf{B}\boldsymbol{a_{0,0}} + \mathbf{n} \quad \textit{target absent},$$

$$\boldsymbol{H_1} \quad : \quad \mathbf{x} = \mathbf{E}\boldsymbol{a_{1,1}} + \mathbf{B}\boldsymbol{a_{1,0}} + \mathbf{n} \quad \textit{target and structured backgroun},$$

where $\mathbf{B}$ is a $d \times Q$ matrix of background materials (the columns of $\mathbf{B}$ are the means of clusters of similar HSI data (background)), and $\mathbf{E}$ is a $d \times P$ matrix of target material spectra (P is the number of columns and, therefore, the variability of the target set). If the target spectra are sparsely located within the image and resolved at sub-pixel levels, then the image data are a mixture of materials from the $\mathbf{Z} = [\mathbf{EB}]$ (which is a concatenated matrix) space. The parameter $a_{0,0}$ is the mixing coefficient (abundance value) of the background material under the null hypothesis, $a_{1,0}$ is the mixing coefficient of the background material under the alternative hypothesis, and $a_{1,1}$ is the mixing coefficient of the target material under the alternative hypothesis. This architecture is depicted in Figure 2.14.

The target detector for this set-up is based on the assumption that the columns of $\mathbf{B}$ and $\mathbf{Z}$ (the means of the clusters for their materials) originate from a multivariate Gaussian distribution

$$p(\mathbf{x}|\mathbf{a}, \mathbf{\Gamma}) = (2\pi)^{-\frac{d}{2}} \left(|\mathbf{\Gamma}|\right)^{-\frac{1}{2}} \exp\left[-\frac{1}{2}\left(\mathbf{x} - \mathbf{Za}\right)^T \mathbf{\Gamma}^{-1} \left(\mathbf{x} - \mathbf{Za}\right)\right], \tag{2.15}$$

where $\left(\mathbf{x} - \mathbf{Za}\right)^T \mathbf{\Gamma}^{-1} \left(\mathbf{x} - \mathbf{Za}\right)$ is the MD of the pixel $\mathbf{x}$ from the cluster mean vector $\mathbf{Za}$. Here, depending on whether the pixel is in the target-plus-background configuration or background only, the $\mathbf{a}$ values take on their appropriate coefficients.

Given this model, the GLRT decides $\boldsymbol{H_1}$ if

$$L(\mathbf{x}) = \frac{p(\mathbf{x}|\hat{\boldsymbol{a}}_\mathbf{1}, \hat{\mathbf{\Gamma}}_\mathbf{1})}{p(\mathbf{x}|\hat{\boldsymbol{a}}_\mathbf{0}, \hat{\mathbf{\Gamma}}_\mathbf{0})} > \gamma.$$

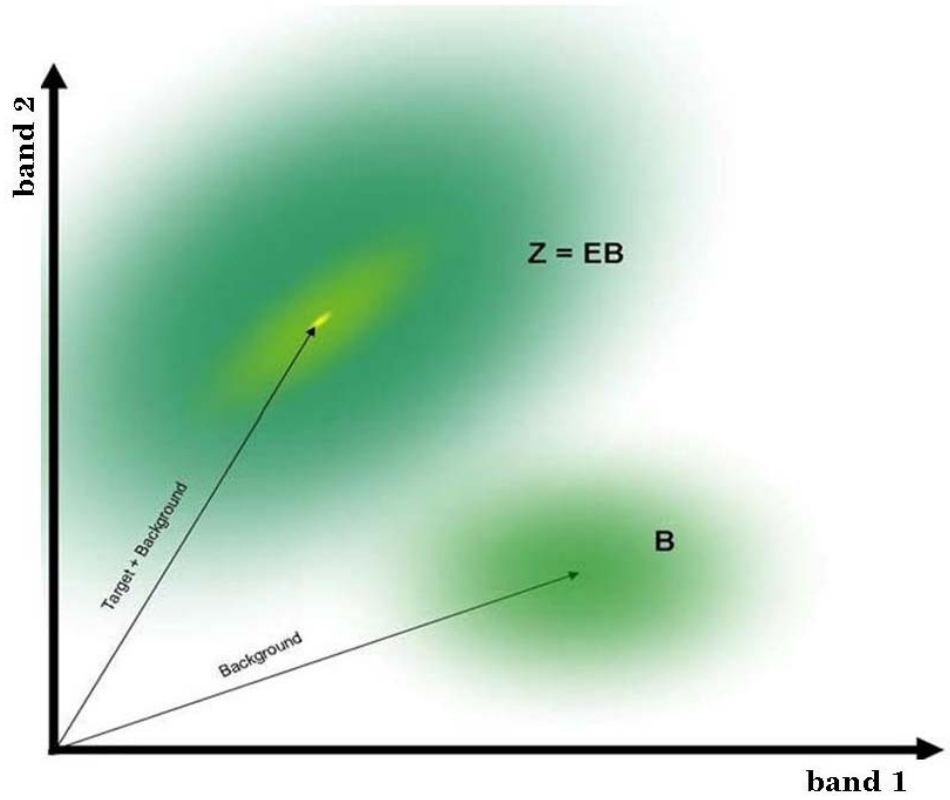


Figure 2.14: The uniform data cloud represents a background pixel variability structure not influenced by the target pixel(s). The data cloud with the inner cloud represents the variability of background experienced by pixels infiltrated by the target data.

The detection statistic for this architecture reduces to [39]

$$D(x) = \left(\hat{\mathbf{\Gamma}}_{\mathbf{0}} - \hat{\mathbf{\Gamma}}_{\mathbf{1}}\right) {\hat{\mathbf{\Gamma}}_{\mathbf{1}}}^{-1}. \tag{2.16}$$

Noting that the maximum likelihood estimate MLE of the covariance matrix is given by

$$\hat{\mathbf{\Gamma}} = \frac{(\mathbf{x} - \mathbf{Za})^T (\mathbf{x} - \mathbf{Za})}{N},$$

the detection statistic is

$$D(\mathbf{x}) = \frac{\mathbf{x}^T \left({P_B}^{\perp} - {P_{EB}}^{\perp} \right) \mathbf{x}}{\mathbf{x}^T {P_B}^{\perp} \mathbf{x}}, \tag{2.17}$$

where ${P_B}^{\perp}$ and ${P_{EB}}^{\perp}$ represent the orthogonal subspace projection operators for the background and target-plus-background, respectively. The distribution of this detector statistic is [39]

$$\begin{aligned} D(\mathbf{x}) &\sim F_{d,N-P} \qquad \textit{under } \boldsymbol{H_0}, \\ D(\mathbf{x}) &\sim F_{d,N-P}(\lambda) \quad \textit{under } \boldsymbol{H_1}, \end{aligned}$$

where $F_{d,N-P}$ is the F-distribution with d and $N-P$ degrees of freedom and $F_{d,N-P}(\lambda)$ is the noncentral F-distribution with non-centrality parameter λ. The fact that the detection statistics are F-distributed is important for the non-normality of HSI data considered in the next section.

The expressions for the distribution of the test statistic relate to the MD distributions evidenced within the HSI data clusters. Expanding (2.17), using the MLEs for the covariances yields

$$D(\mathbf{x}) = \frac{(\mathbf{x} - \mathbf{Ba})^T {\boldsymbol{\Gamma_1}}^{-1} (\mathbf{x} - \mathbf{Ba})}{N} - \frac{(x - Za)^T {\boldsymbol{\Gamma_1}}^{-1} (\mathbf{x} - \mathbf{Za})}{N}, \tag{2.18}$$

which is a scaled difference between the MD of pixels containing only background in the cluster of target-plus-background and the MD of pixels containing target and background in the same cluster. Therefore, $D(\mathbf{x})$ exploits the information in the MDs evidenced by the data. This scenario is depicted in Figure 2.15.

Equation (2.17) is also the CFAR Matched Subspace Filter for non-coherent detection as defined by Scharf [44] and is the fourth and final in the suite of adaptive

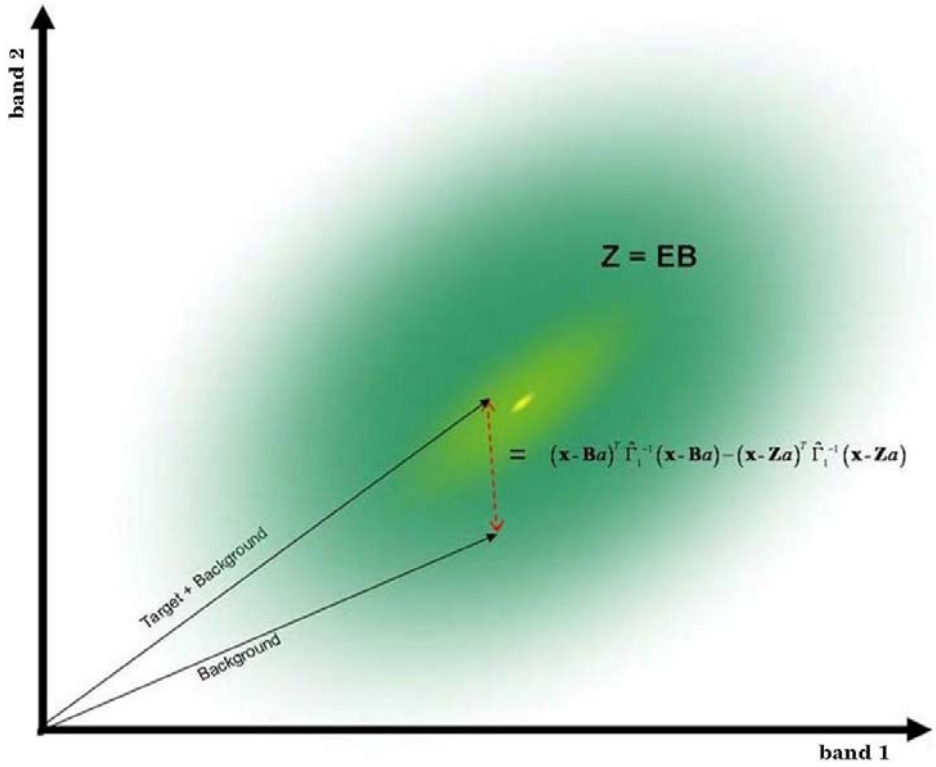


Figure 2.15: Detector for the structured background hyperspectral model. The dashed line in the data cloud represents the MD between a pixel containing target and background and a pixel containing only background.

statistical target detectors common in exploiting signals in subspace architectures. The detectors presented here overview the exploitation tools for the assumed models. A more detailed taxonomy of the different types of detectors for HSI data exploitation is given by Keshava, et al. [41], Manolakis, et al, [53], and Stein, et al. [78].

Having established the significance of MDs in the over-arching framework for HSI signal processing, the next section describes the effect of MD distributions on the metrics obtained from the detectors. Models of MD distributions from HSI data clusters are then considered. Next, the latest work, including work from this dissertation, in modeling MD distributions is introduced.

2.3 MD Effects on Metrics

In addition to the design of a target detection algorithm, its performance must be measured. A common metric for target detection is the Receiver Operating Characteristic (ROC) curve. A brief description of ROC curves associated with the first two target detectors introduced in the previous section is given here.

For the matched filter (under the full-pixel model), since $\mathbf{x}$ is assumed normal under both hypotheses and the log-likelihood ratio of the hypotheses is a linear transformation of the variable $\mathbf{x}$, the detection statistic is [70]

$$\begin{aligned} y &\sim N\left(0, \Delta^2\right) \text{ under } \boldsymbol{H_0} \\ y &\sim N\left(\Delta^2, \Delta^2\right) \text{ under } \boldsymbol{H_1}, \end{aligned}$$

where Δ^2 represents the squared Mahalanobis distance (MD). The probability of detection and the probability of false alarm are then calculated using the above distributions under each hypothesis. The ROC curve for the matched filter is given in Figure 2.16.

Obviously, from Figure 2.16 and from Equation (2.8), as the test pixel resides closer to the target in the HSI space (i.e., larger Δ^2 from the test pixel to the background space), the detector performance increases. For the anomaly detector of Figure 2.12, the detector statistic is

$$y \sim (\mathbf{x} - \boldsymbol{\mu_0})^T \boldsymbol{\Gamma}^{-1} (\mathbf{x} - \boldsymbol{\mu_0}) . \tag{2.19}$$

Since the expression is quadratic, the detector is distributed as a non-central chi-squared distribution. The distribution under the two hypotheses is [49]

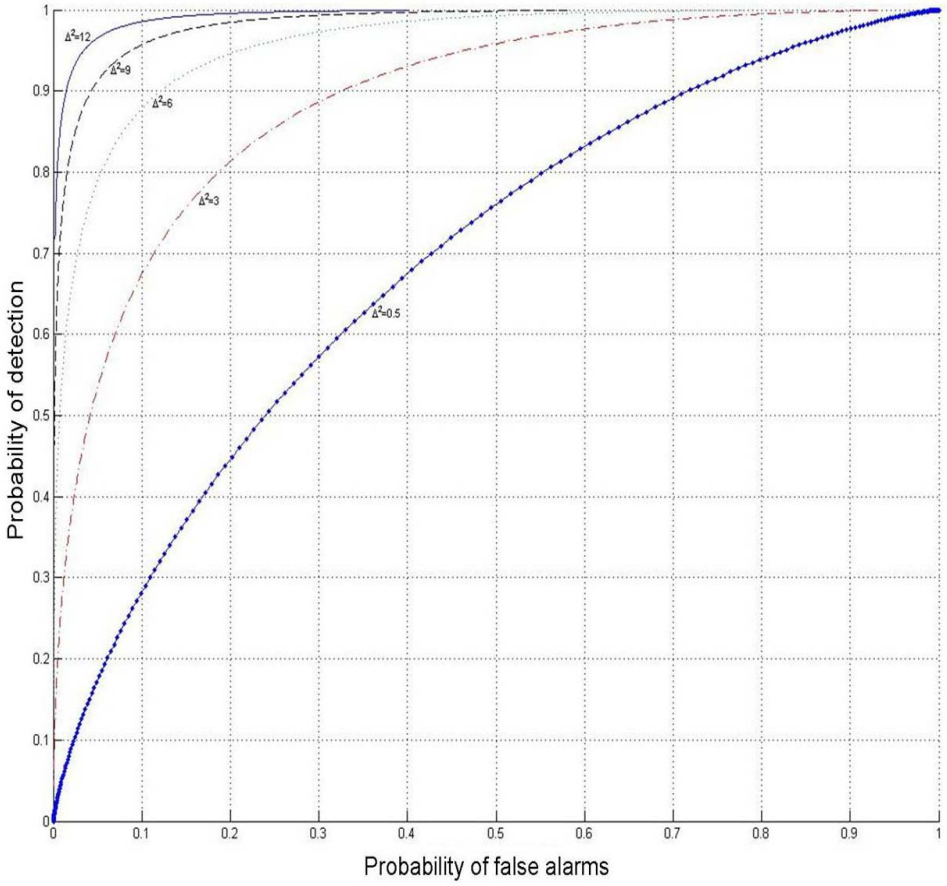


Figure 2.16: ROC curves for the matched filter target detector. ROC curves for different Mahalanobis distance (Δ^2) values are given.

$$
\begin{aligned}
y &\sim \chi_d^2(0) \; under \; \boldsymbol{H_0} \\
y &\sim \chi_d^2(\Delta^2) \; under \; \boldsymbol{H_1},
\end{aligned}
$$

where $\chi_d^2(\rho)$ is a non-central chi-squared distribution with d degrees of freedom (corresponding to the dimensionality of the data) and non-centrality parameter ρ. The probability of detection and the probability of false alarm are then calculated using the above distributions under each hypothesis. The ROC curve for the anomaly detector is shown in Figure 2.17.

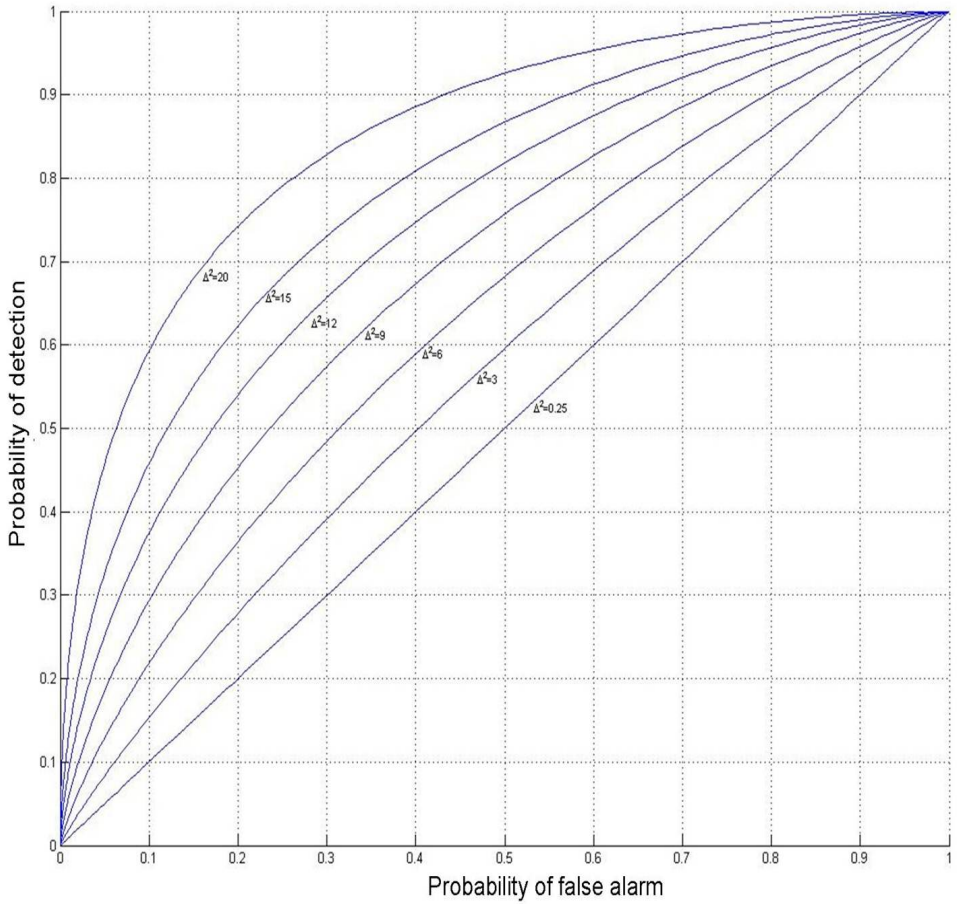


Figure 2.17: ROC curves for the anomaly detector. The ROC curves for different Mahalanobis distance (Δ^2) values are shown.

The detection statistic for the unstructured background anomaly detector is distributed as a student-t, and for structured backgrounds it is distributed as an F-distribution [43, 44, 49]. Therefore, the performance of the detector depends on the distribution of the detection statistic. Proper modeling of the detection statistic leads to optimized performance for the target detector.

2.4 Distribution Model for Hyperspectral Data

In order to achieve the correct distribution for the detection statistic, the distribution of the data model under each hypothesis must be correct. In each of the detectors given in the previous section, the HSI data model is assumed to be Gaussian.

However, previous work [46, 54, 61] has shown that the underlying data distribution for HSI is not Gaussian.

Recent papers on this topic show that the data tend to follow a family of elliptically-contoured EC distributions [33,48,51,56], of which the normal distribution is a small subset. Specifically, MD data tend to follow univariate heavy-tailed distributions, which imply multivariate heavy-tailed EC distributions. An EC probability density function is characterized by [8, 21]

$$p(x|\Sigma) = c\,(|\Sigma|)^{-\frac{1}{2}}\,h(\Delta^2), \tag{2.20}$$

where Σ is a scale matrix, c is a normalizing constant, and $h(\Delta^2)$ is a function of Mahalanobis distance. The normal distribution is a special case of this family for which $h(\Delta^2) = e^{-\frac{\Delta^2}{2}}$.

The distribution of the MDs may be used to describe a multivariate data set. For example, MDs are distributed as univariate chi-squared with d degrees of freedom for a population of d-dimensional multivariate normal random data vectors [37]. Thus, the normality of a multivariate data set can be tested by observing the behavior of its MD distribution [67]. Specifically for HSI data, by observing the distribution of MDs from clusters of similar pixel spectra, the underlying distribution of scene variability may be determined. The distribution of cluster data is fundamental for analyzing variability in the entire HSI scene.

In Equation (2.3), a single pixel is modeled as resulting from a mixture of pure spectra from pixels originating in hypothetical material class distributions. This equation can be interpreted as

$$f_d\left(\mathbf{x}\right) = \sum_{k=1}^{N} a_k f_d\left(\mathbf{x} \mid \Theta_k\right), \tag{2.21}$$

where $f_d\left(\mathbf{x} \mid \Theta_k\right)$ is a single unimodal d-dimensional multivariate distribution for cluster k, a_k is the probability that cluster k is a constituent of the scene, also known as

the abundance value, N is the total number of clusters defined by the materials in the scene, and Θ_k is a vector of parameters that describe the distribution associated with cluster k.

The parameters Θ_k may be readily determined by implementing any of a number of parameter estimation routines. However, the expectation maximization method and its variant, stochastic expectation maximization (SEM), are the more robust routines applied for HSI [18, 54]. Once the parameters are estimated, the individual clusters describing the separate classes of materials in the image are developed.

The left image in Figure 2.18 shows a small HSI subsection. The image data is clustered into material classes that display similar properties using the SEM algorithm. These properties are parameterized by the means $\boldsymbol{\mu_k}$ and covariances $\boldsymbol{\Gamma_k}$ which describe the distribution of the material class within each cluster. The right image in Figure 2.18 is a two-dimensional representation of data clustering in the d-dimensional space in which the image is classified, with each color representing a different class of material (i.e., a unique cluster).

Upon identifying the clusters, it can be seen (particularly in the classification image) that there is some variability within each cluster. For example, for the tarmac cluster in the left image of Figure 2.18 (largest class), three of the building tops are grouped into the tarmac cluster. This variability within clusters leads to the different shapes apparent in each MD distribution for each cluster.

Nominal procedures in hyperspectral imaging for remote earth sensing dictate that large land cover areas be encompassed in the field of view. This requirement results in many different material types as constituents of the data set. The variability of the different materials is collectively modeled as a multi-modal multivariate distribution. Separating each material into its respective class yields multivariate cluster distributions that exhibit unimodal multivariate shapes, where "class" refers to material type and "cluster" refers to collections of samples of a material type [54, 77]. Figures 2.19 - 2.22, created using ENVI® [76], illustrate these concepts. Each cluster

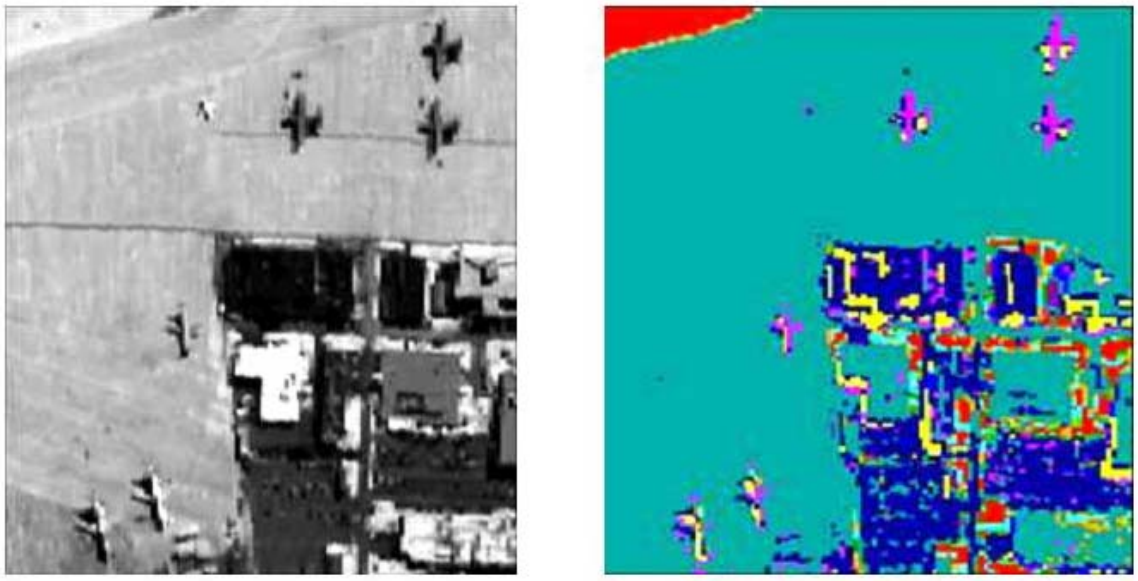


Figure 2.18: Left: Grayscale image of Harrisburg International Airport taken by the AVIRIS [24, 45] HSI sensor (left top quadrant = grass, right lower quadrant = tops of buildings, right top quadrant = three airplanes on tarmac, left lower quadrant = three different airplanes on tarmac). Right: Image showing the different materials found by the clustering routine. Note that some building tops are grouped into the tarmac cluster.

of data results in a unique MD distribution for the data set. Some examples of the distributions are given in Figure 2.23.

2.5 Fitting MD Distributions

Figure 2.24 shows a typical histogram of MDs from a HSI cluster and a plot of a univariate chi-squared distribution. Notice that the chi-squared distribution poorly models the shape of the data, especially in the tail region. This result implies that the underlying distribution of the cluster data set, which creates the MD distribution, is not Gaussian. The distribution must be determined to generate a correct statistical model for developing detection hypotheses.

In [54] it is shown that a mixture of F-distributions models the MD distribution data more accurately. Specifically, it is demonstrated that a mixture of two F-distributions, one modeling the body of the data and another incorporating infor-

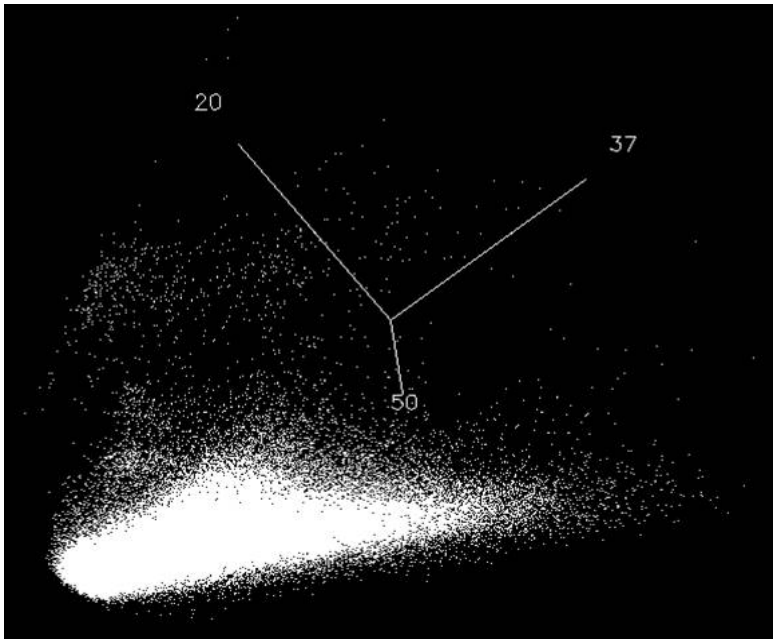


Figure 2.19: HSI data points (250,000) from an image projected onto bands 20, 37, and 50 out of a total of 224 bands. The points are the resultant of the vector created by the coordinates in three axes (pixel intensity at the given band). The result is a three-dimensional representation of the multi-modal 224-dimensional distribution of the data.

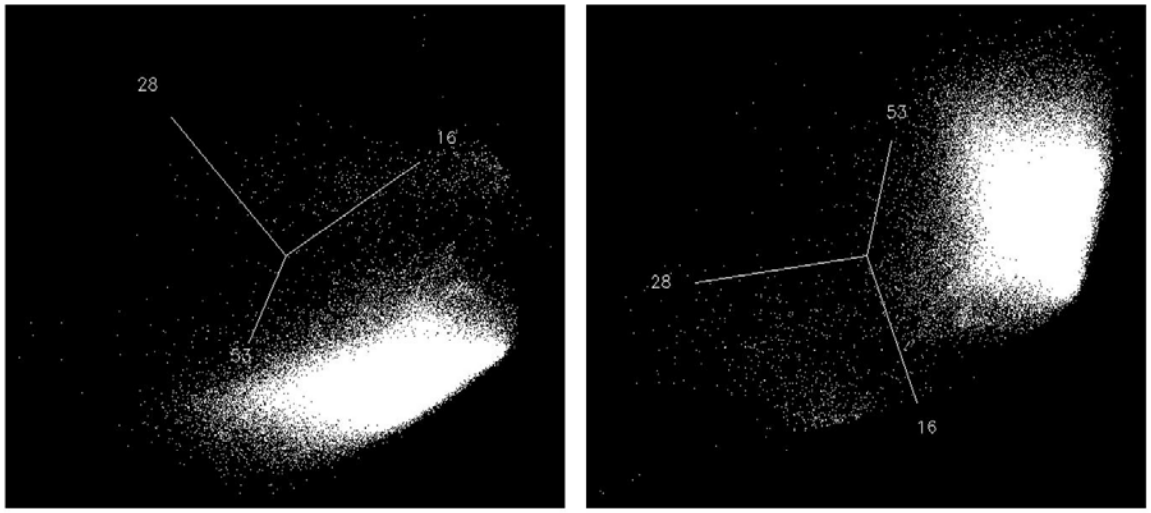


Figure 2.20: The same 250,000 HSI data points plotted against three different coordinates (bands). Notice the changing shape of the data cloud and the concentration of data points in certain regions as the axes are rotated. These concentrations are clusters of similar data.

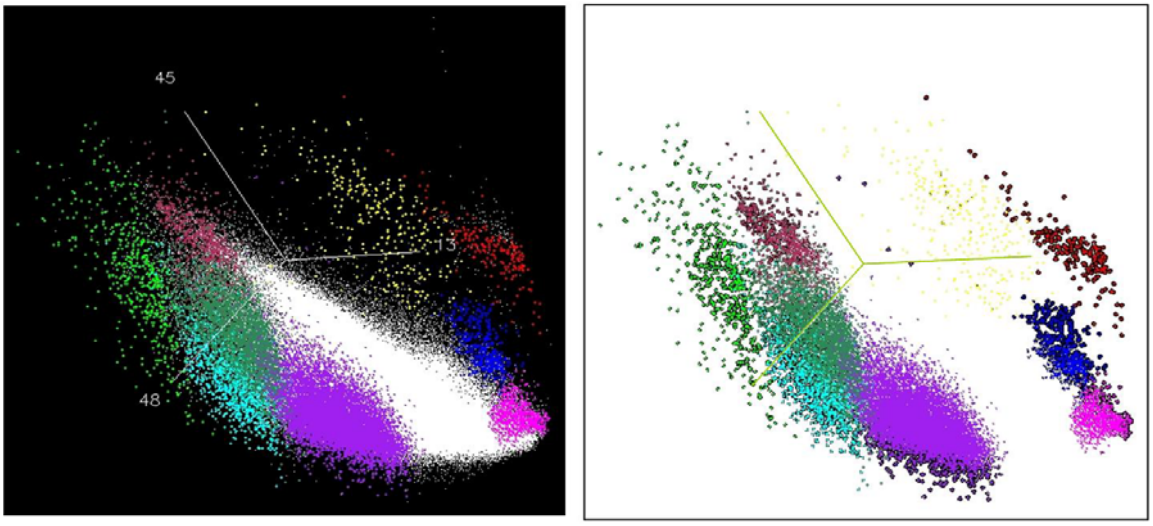


Figure 2.21: The same 250,000 HSI data points plotted against another three different coordinates. The results show that the data cloud is different when examined in different dimensions. Also plotted are 8 clusters of data (clustered using the K-means algorithm). The figure on the right shows the clusters only (with the un-clustered data left out).

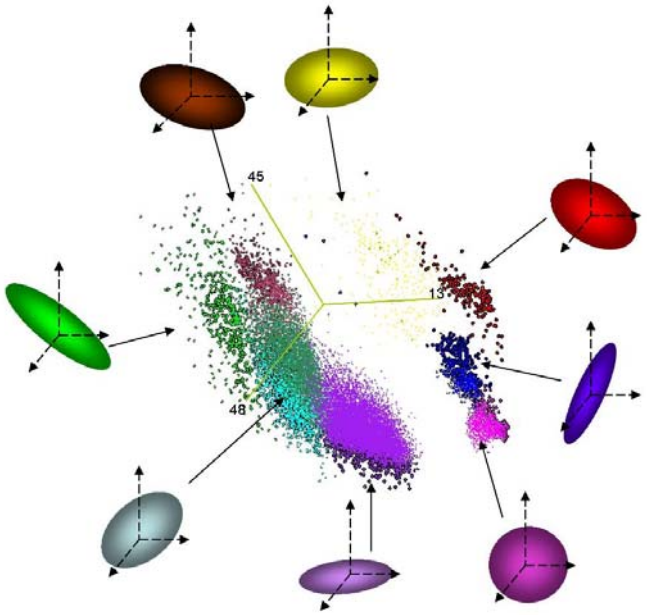


Figure 2.22: The same clusters shown in the previous figure. Here each ellipsoid represents the material cluster distribution from which the majority of the spectral information for the corresponding clustered pixels originates (assuming that each pixel is a full-pixel representation of only one material).

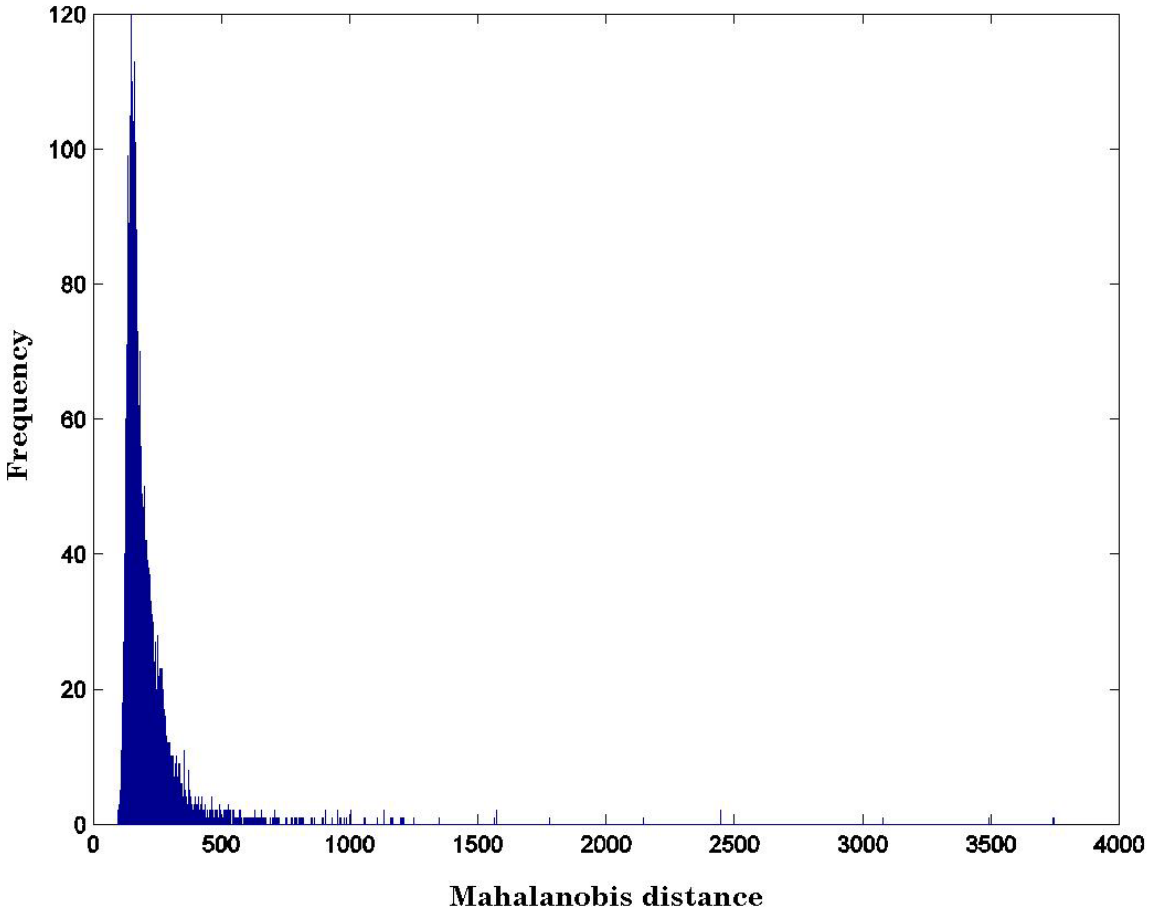


Figure 2.23: Distribution of a cluster of data taken from the 250,000 hyperspectral data points given above (cluster determined using the K-means algorithm). Notice the long tail of the distribution.

mation from the tails of the data, creates an accurate overall model. An example of this derived distribution for the above data set is shown in Figure 2.25.

Visual inspection indicates a better fit to the data. However, here a quantitative test is performed to better assess the fit. As in [54], plots of the probability of exceedance of the data and the fitted distribution are generated. The probability of exceedance is the probability of exceeding a given value in the range of a distribution and is $Q(y) = 1-G(y)$, where G is the cumulative distribution function. For example, the probability of exceedance for $y = 0$ when $f(y)$ is a normal distribution with zero mean and unit variance is 0.5, since half of the distribution lies on one side of $y = 0$, and as y increases from zero the probability of exceedance decreases. This decrease

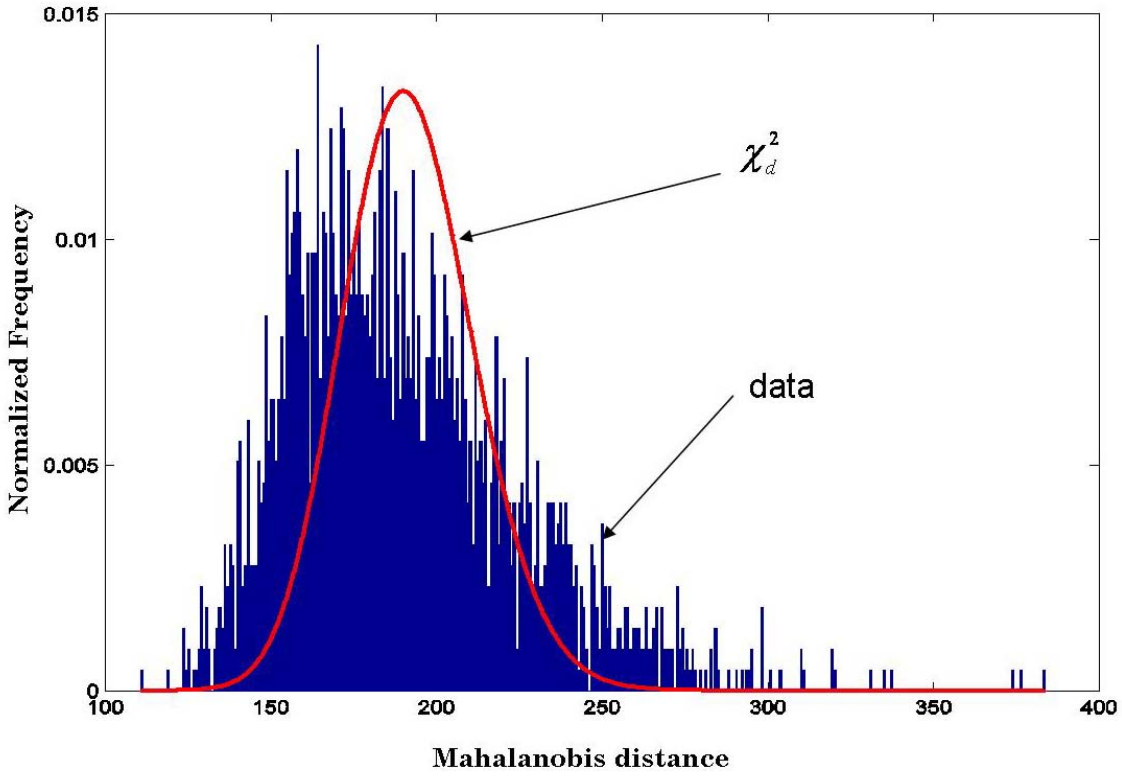


Figure 2.24: Histogram of Mahalanobis distances from a cluster taken from the 250,000 hyperspectral data points given above and a Chi-squared distribution fit (smooth curve). Notice the poor fit in the tail of the distribution.

is due to the decreasing portion of the distribution (the tail) as y increases. The probability of exceedance is a suitable metric for comparing tails of distributions, as varying tail widths generate different probability of exceedance values.

Probability of exceedance is plotted in Figure 2.26 for the MD data of Figure 2.25 and for the fitted distributions. A method for comparing these curves is the mean squared error MSE between the probability of exceedance curve for the fitted distribution and the probability of exceedance curve of the data. This metric is used because probability of exceedance incorporates the "tails" of the data in determining the shape of the model (more so than other model fit metrics, such as chi-squared). In [54] it is shown that it is in the shape of the "tails" that HSI data modeling is most important.

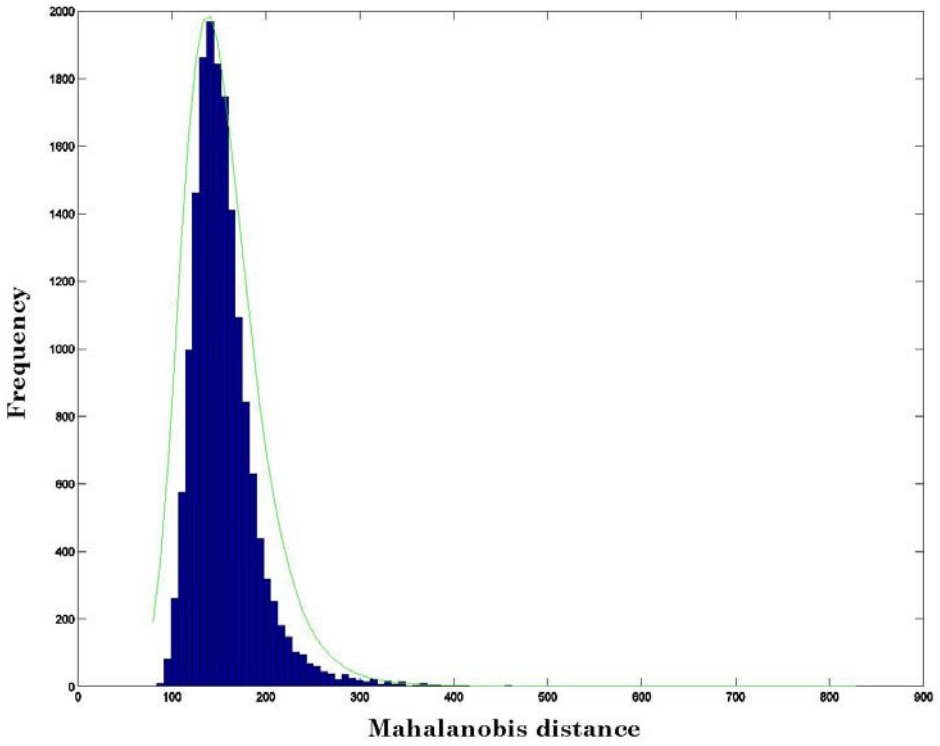


Figure 2.25: Histogram of 17,453 MDs and a mixture of two F-distributions (smooth curve). The fit in the tail region is improved.

Figure 2.26 shows that the mixture of F-distributions models the data better than the chi-squared distribution. The mixture of F-distributions is developed by weighting an F-distribution that models the body of the MDs along with an F-distribution which models the behavior of the tails [54]. F-distributions are a member of the heavy-tailed univariate distribution family [36], and thus the MD data may be modeled by other heavy-tailed distributions. A mixture of F-distributions that models the MD distribution of a cluster implies that the underlying distribution describing the statistics of the HSI data set is a mixture of multivariate t-distributions [58,74,79]. The multivariate t-distribution is

$$x_t \sim \frac{\Gamma(\frac{d+\nu}{2})}{(\pi\nu)^{\frac{d}{2}}\Gamma(\frac{\nu}{2})}|\Sigma|^{-\frac{1}{2}}\left(1+\frac{\Delta^2}{\nu}\right)^{-\frac{d+\nu}{2}}. \tag{2.22}$$

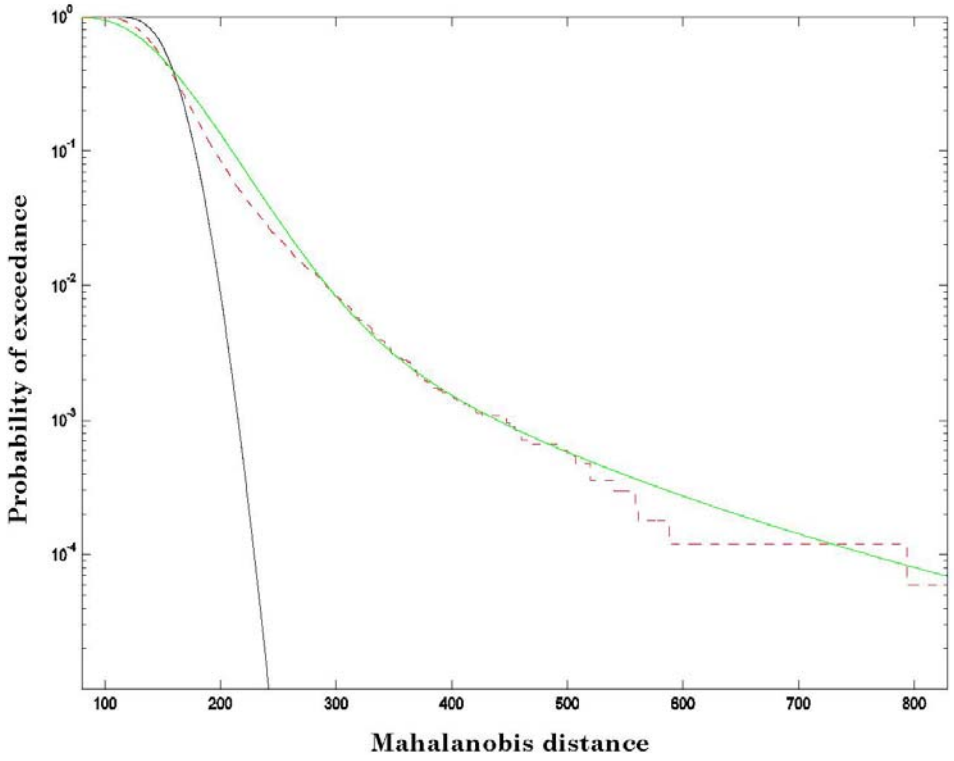


Figure 2.26: Exceedance plot for the MD distribution of a cluster of HSI data (dashed line), Chi-squared distribution (dark solid curve) and F-mixture distribution (light solid curve). The Chi-squared plot is obtained if the MDs are distributed normally. Notice the heavier tail exhibited by the MD distribution.

where $\Gamma(\xi)$ is the gamma function with argument ξ, Σ is a scale matrix, ν is the degree of freedom, and Δ^2 is the MD term. The multivariate t-distribution is a member of the multivariate EC family (in agreement with the statements made in first part of Section 2.4). If $\nu = 1$, the distribution is multivariate elliptical Cauchy, and if $\nu \to \infty$, the distribution is multivariate Gaussian [79].

The multidimensional distribution of a hyperspectral image data set is therefore modeled as

$$t_d\left(\mathbf{x} \mid \Theta\right) = w \cdot t_d\left(\mathbf{x} \mid \Theta_1\right) + (1 - w) \cdot t_d\left(\mathbf{x} \mid \Theta_2\right), \tag{2.23}$$

where w is a mixing coefficient such that $0 \leq w \leq 1$ and Θ_i are the respective parameter sets for each t-distribution. The MD distribution model is then

$$MD \sim w \cdot F_{d,\nu_1} + (1 - w) \cdot F_{d,\nu_2}. \tag{2.24}$$

Here the body of the MD distribution is modeled by the first term in Equation (2.24) and the tail is modeled by the second term.

Manolakis and Marden [54] show that by modeling the data with multivariate t-distributions, the performance of the adaptive anomaly detector improves. By modeling the MD distributions with a mixture of F-distributions, the detector threshold is more accurately selected and considerably lower false alarm rates are obtained (compared to modeling under the multivariate normal assumption). Also, robust synthetic data with the same distribution as HSI data is generated by using a mixture of multivariate t-distributed random d-dimensional vectors. Such simulated data (previously unattainable) are valuable for evaluating detection and classification algorithms.

The significance of MDs in the stochastic model for HSI signal processing is developed, and their effects on the metrics associated with the output of these processes is demonstrated. It is also shown how variability in HSI data causes varying MD distribution shape. This shape is not properly modeled by Gaussian assumptions for the HSI data set, which result in inaccurate models for the MDs and, therefore, processing limitations.

By fitting MD distributions with a mixture of F-distributions, a more accurate model of the statistics is obtained. Specifically, F-distributed MDs result in multivariate t-distributed data, a heavy-tailed member of the EC distribution family. This result implies that heavy-tailed distributions are suitable for modeling MD distributions and, hence, for obtaining a robust stochastic model for the HSI data set.

A limitation of using a mixture of F-distributions to model the MDs is the lack of computationally efficient methods to solve for the parameters ν_1 and ν_2. Current methods involving bi-variate exhaustive search techniques and maximum likelihood estimation methods are computationally inefficient and often result in sub-optimal solutions [54, 58, 74]. Therefore, there is a need to develop automated, robust, and

computationally efficient methods for modeling MD distributions. The next chapter addresses this goal.

2.6 Research Roadmap

This section provides a brief methodology roadmap and lists the objectives for the present research given the above described phenomena. The first step identifies which families of distributions include members with parameter values that result in heavy-tailed distribution behavior (as MD distributions exhibit heavy-tails). Initial research shows that the Johnson family of distributions contains a subset that describes heavy-tailed behavior and provides a method for estimating parameters efficiently from the data set [35, 75]. Also, there is vast literature in the fields of econometrics and actuarial sciences in extreme value statistics which suggests the use of Pareto models to fit heavy-tailed behavior [3, 9, 10, 12]. An investigation by Manolakis [52] points to the feasibility of extreme value statistical methods in HSI modeling. Therefore, these two areas are investigated as described below.

Objective 1: Determine an optimal Johnson distribution model for HSI MD data. The Johnson distribution system covers a wide range of different distribution types. An initial investigation into fitting MD distributions with the Johnson system is performed. Once a specific Johnson distribution type is identified, the model is optimized for robustness against possible outliers and perturbations in data sets, and is automated for maximum computational efficiency. This model is then compared to the F-distribution mixture model.

Objective 2: Determine a multivariate elliptically contoured distribution model from the univariate Johnson distribution modeling the MDs. A multivariate EC model is derived from the univariate Johnson distribution modelling the MD data. This multivariate model is developed using the definitions described by EC theory. The model is compared to derivation of the multivariate t-distributed EC model from the univariate F-distribution of MDs.

Objective 3: Determine a viable parameter estimation method for obtaining tail-index parameters for GPD models. Different parameter estimation methods are analyzed for fitting a generalized Pareto distribution (GPD) to MD data. The different methods are tested with respect to suitability for HSI MD data models, and to sensitivity of the estimators to smaller subsets of an entire MD data set. The most robust estimation method is determined and most effective data subset threshold is established. A smaller subset of the entire MD data set may be used for greater computational efficiency.

Objective 4: Develop an optimal method for obtaining robust tail-index estimates for GPD models of MD distributions. Once an efficient estimation method for GPD models for MD distributions is identified, the method is optimized to generate minimum MSE with respect to possible outliers and perturbations in data. An automated, robust, and computationally efficient algorithm is developed for the identified GPD parameter estimation method most suitable for HSI data processing.

Objective 5: Assess the utility of Johnson and GPD models for stochastic HSI data processing. The Johnson model is compared against the GPD model, and the estimation methods and their robustness under different data configurations are assessed. The method which yields better results and provides more information about the process is then determined.

The chapters that describe the research involved in achieving the objectives are as follows:

Chapter III: initiates the research with the goals described in Objectives 1 and 2. A specific Johnson distribution is selected based on properties similar to MD data behavior. Fitting the distribution to MD data is optimized with respect to potential outliers (secondary processes) which create irregularities in forecasted data behavior. A new metric for gauging goodness-of-fit is developed and applied to the new model.

Chapter IV: objective 3 is addresses with results and initial conclusions on the method developed. Specifically, the theory of extreme values is reviewed and

threshold methods are applied to HSI data with a number of corresponding parameter estimation techniques. The families of estimation techniques are evaluated and an optimal routine is identified for fitting GPDs to HSI MD data. The routine is analyzed with respect to threshold sensitivity and an optimal threshold level is obtained.

Chapter V: the estimation method identified in the previous chapter is optimized for robustness against noised and possible outliers in the data. A two-pass algorithm is developed which uses a feedback mechanism to adjust estimated parameters based on a predetermined level of irregularity in the data model. The robust estimation method is applied to HSI MD data to verify performance enhancement.

Chapter VI: all information from the previous chapters is used to assess the utility of using the Johnson distribution and extreme value methods, and a final assessment is reported on each method as described in Objective 5.

Chapter VII: final conclusions, a summary of all research is compiled, and suggestions for further research are provided.

III. Johnson System Models of MD Distributions

Of the five objectives outlined in the previous chapter, the first two deal with identifying a Johnson distribution for modelling heavy-tailed MD distributions. This chapter analyzes the Johnson distribution system and its application to modelling MD distributions. Specifically, the most efficient and robust form of a Johnson distribution for the task is analyzed and applied.

3.1 Johnson Distributions

3.1.1 Background. The Johnson system estimates an empirical distribution by applying a transformation to a standard normal distribution [28, 75]. The three types of Johnson distributions, S_U, S_L and S_B, cover a wide range of distribution families. In the univariate case the S_U distribution is

$$z(x|\gamma_J, \eta_J, \epsilon_J, \lambda_J) = \gamma_J + \eta_J \sinh^{-1} \frac{x - \epsilon_J}{\lambda_J}, \quad -\infty < x < \infty, \tag{3.1}$$

where z is a standard normal variate, ϵ_J and γ_J are location variables, and η_J and λ_J affect scale and shape. The S_U distributions include the Gaussian distribution, t-distribution, Cauchy distribution, and other families of unbounded distributions where the support is from negative to positive infinity. The S_L distribution is

$$z(x|\gamma_J, \eta_J, \epsilon_J, \lambda_J) = \gamma_J + \eta_J \ln \frac{x - \epsilon_J}{\lambda_J}, \quad x \geq \epsilon_J, \tag{3.2}$$

which covers the F-distribution, χ^2 distribution, and others where the logarithm of the random variable follows the normal distribution. Finally, the S_B distribution is

$$z(x|\gamma_J, \eta_J, \epsilon_J, \lambda_J) = \gamma_J + \eta_J \ln \frac{x - \epsilon_J}{\lambda_J + \epsilon_J - x}, \quad \epsilon_J \leq x \leq \epsilon_J + \lambda_J, \tag{3.3}$$

which includes distributions bounded by $(\epsilon_J, \epsilon_J + \lambda_J)$, such as the beta and uniform distributions.

3.1.2 Mechanics. The Johnson parameters are obtained by applying an autonomous algorithm developed by Slifker and Shapiro [75]. This algorithm is based on selecting four equally spaced percentiles (quantiles) from the standard normal distribution indicated by $-3z$, $-z$, z, and $3z$. Usually, these quantiles are selected at $z = 1$, $-z = -1$, $3z = 3$, and $-3z = -3$, corresponding to normal distribution percentiles of 84.13%, 15.87%, 99.87%, and 0.13%, respectively. These percentiles are then determined for quantile locations on an unknown distribution represented by x_{-3z}, x_{-z}, x_z, x_{3z}.

Slifker and Shapiro demonstrate that [75]

$$\begin{aligned} \frac{mn}{p^2} &> 1, for\ an\ S_U\ distribution, \\ \frac{mn}{p^2} &< 1, for\ an\ S_B\ distribution, \\ \frac{mn}{p^2} &= 1, for\ an\ S_L\ distribution. \end{aligned}$$

where $m = x_{-3z} - x_z,\ n = x_{-z} - x_{-3z},\ and\ p = x_z - x_{-z}$. Using these criteria, a candidate Johnson transformation (one of Equations (3.1), (3.2), or (3.3)) is selected to fit the unknown distribution by transforming a normal distribution accordingly.

For this research, only the Johnson S_L distribution is used. Since MD distributions are positive (bounded at $MD = 0$) and un-bounded toward $+\infty$, the only applicable distribution is the Johnson S_L. The other two are either un-bounded in the positive and negative direction (S_U), or bounded on both sides (S_B).

3.1.3 Simulations. In what follows, different distribution forms are simulated by generating data from a set of known distributions. In particular, 10,000 MD values are generated randomly for specified parameters of a given distribution, and the parameters of the Johnson S_L distribution which result in a minimum squared

error fit to the data are found. An exceedance log-plot of the data and the Johnson S_L distribution fitting the data are given to show the fit at the tail of the data more clearly.

The first simulated distribution is a gamma distribution, which in the Pearson system is a type III distribution, of which the χ^2 distribution is a special case. The univariate χ^2 distribution of MD is obtained for a multivariate Gaussian distribution of the data (the general assumption about HSI data until recently). A Johnson S_L distribution fit to the gamma distributed data, where the gamma distribution parameters are shape = 10 and location = 155, is shown in Figure 3.1.

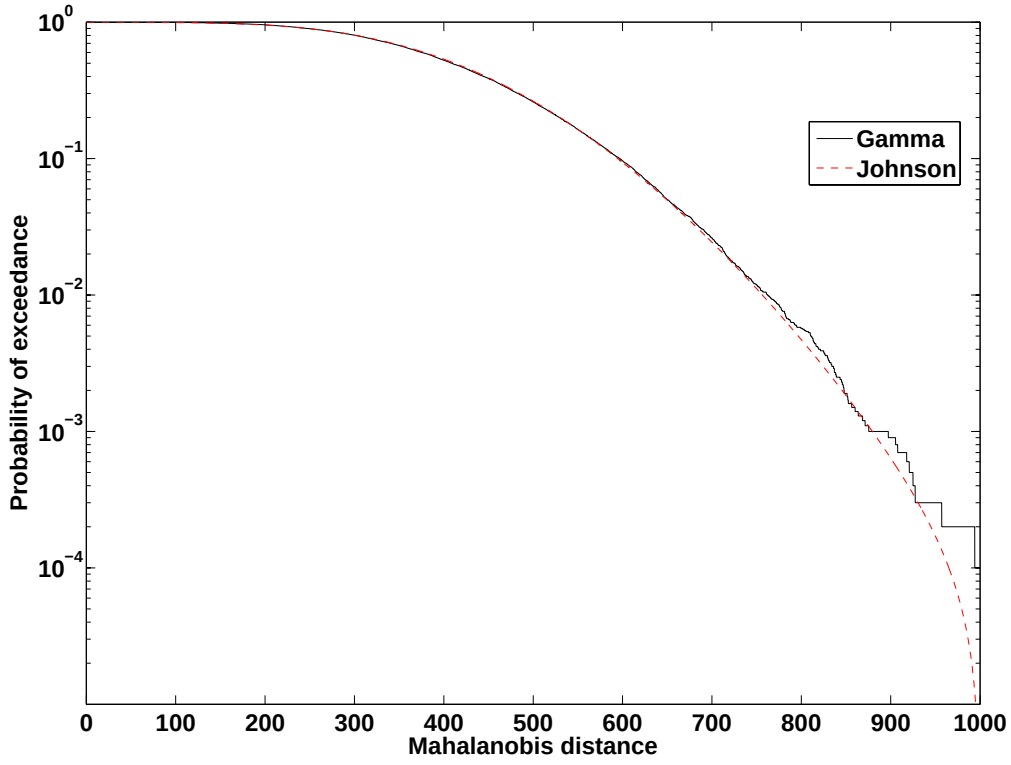


Figure 3.1: Johnson S_L distribution (dotted line) fit to 10,000 values generated from a Gamma distribution with parameters: shape = 10 and location = 155 (solid line).

Figures 3.2, 3.3, and 3.4 depict Johnson distributions fit to Weibull, Lognormal and F-distributed data, respectively, where the parameters used to generate the plots

are given in the Figure caption. These distributions are selected because of their heavy tails (tails that decrease more slowly than asymptotically exponential). It is well-known that HSI MD data exhibits this heavy-tailed property. Figure 3.5 depicts a Johnson S_L distribution fit to data from a mixture of two F-distributions, which is the model recently used to model the variability in HSI MD distributions [54]

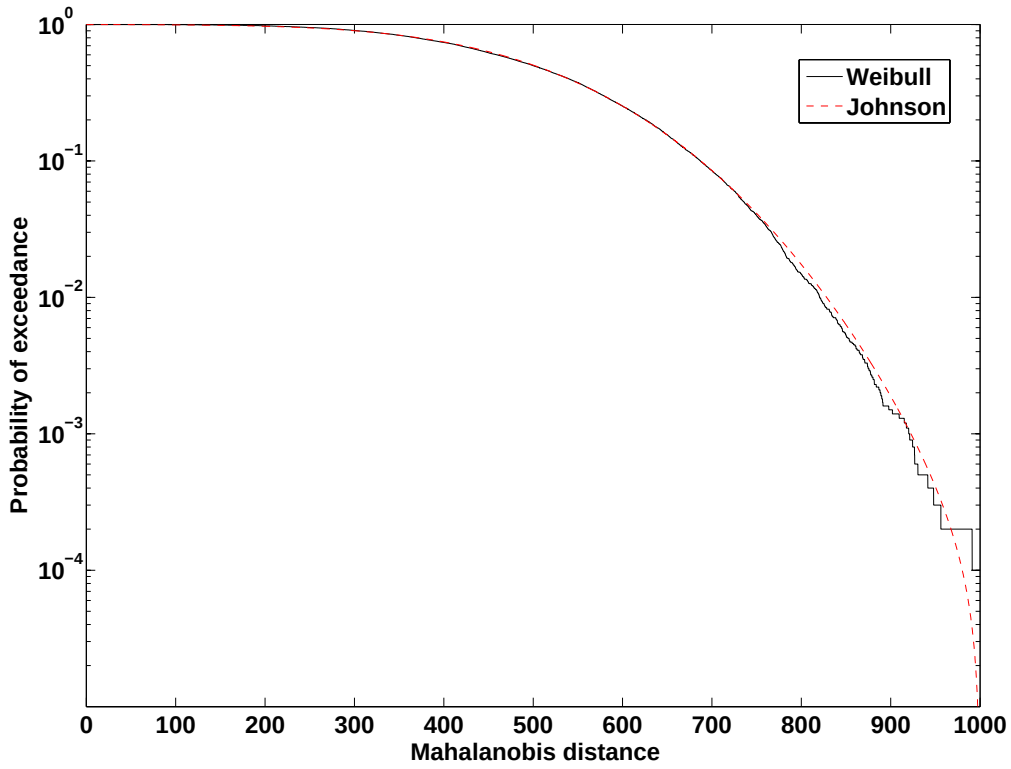


Figure 3.2: Johnson S_L distribution (dotted line) fit to 10,000 values generated from a Weibull distribution with parameters: scale = 25 and shape = 10 (solid line).

3.1.4 Motivation. From Figures 3.2 - 3.5, notice that the Johnson S_L distribution fits all of the simulated data well, which reflects the type of distribution and, therefore, the amount of information in the tails with respect to the body. Graphically, this phenomenon can be traced to regions on the β_1, β_2 chart [79] shown in Figure 3.6. Parameter β_1 is a measure of skewness (i.e., if $\beta_1 = 0$, a distribution is symmetric, and increasing values of β_1 denote increased positive skew) and β_2 is a

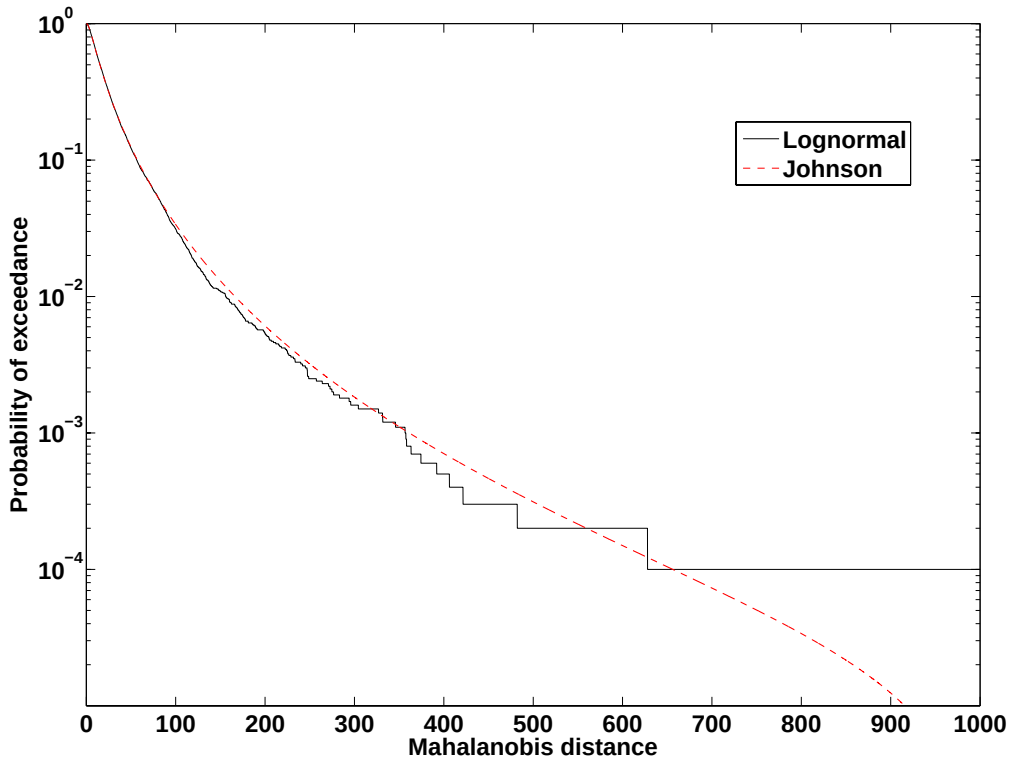


Figure 3.3: Johnson S_L distribution (dotted line) fit to 10,000 values generated from a Lognormal distribution with parameters: mean = 155 and variance = 2 (solid line).

measure of kurtosis (larger values for flatter distributions and smaller values for a higher degree of peakedness). The range of distributions fit by the S_L distribution (depending on threshold width) is defined in Figure 3.6. With $mn/p^2 = 1$ the S_L distribution is confined to the lognormal line. However, in practice, the S_L criterion is given a bandwidth $mn/p^2 = 1 \pm w$, where w is a constant that allows the S_L distribution to define a region above and below the lognormal line.

Here it is determined that $w = 0.3$ above the lognormal line, and $w = 0.6$ below the lognormal line, which allows the S_L distribution to fit the range of Pearson Type VI distributions (F-distribution, Weibull forms and lognormal distribution), as well as those approaching Pearson Type III (Gamma distributions and another Weibull form). Also, this choice enables fitting to a wide range of Type IV distributions (some

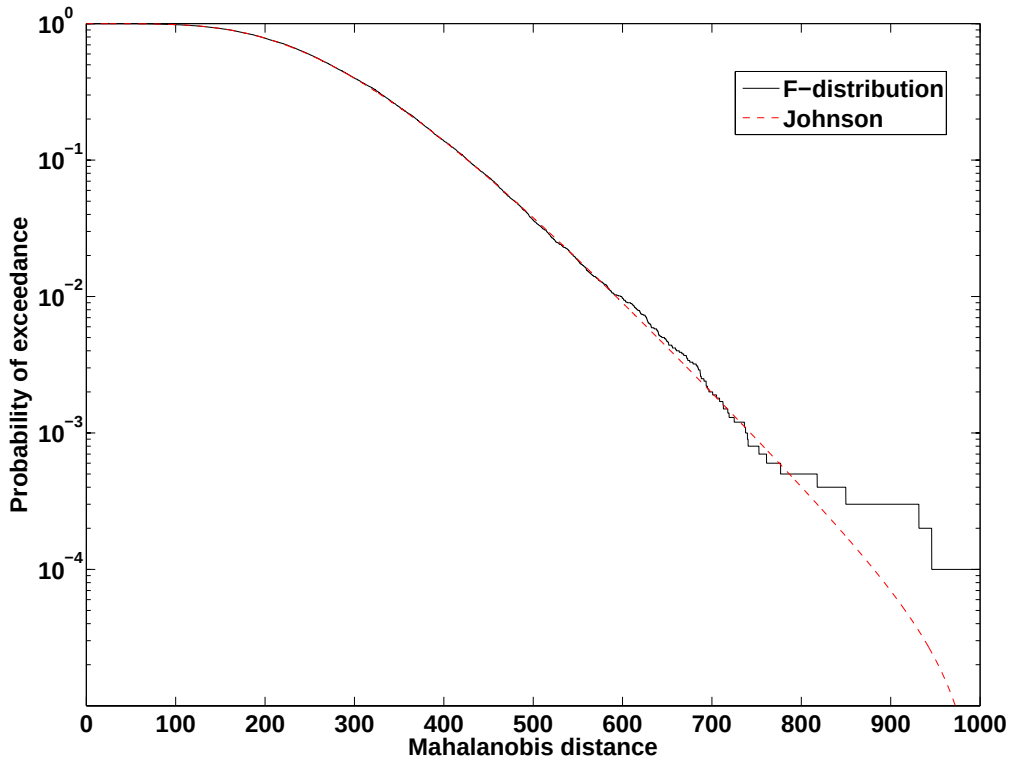


Figure 3.4: Johnson S_L distribution (dotted line) fit to 10,000 values generated from a F-distribution with parameters: $\nu_1 = 155$ and $\nu_2 = 100$ (solid line).

Bessel distributions and other exotic distributions that lack a closed form, which have infinite support in both directions but are shifted such that values approaching zero and less than zero are extremely small, are unimodal, and take on various right-tail weights).

Based on the research conducted here, the region in Figure 3.6 just below the Pearson Type III line and to the right of $\beta_1 = 1$ locates the majority of HSI MD distributions. This region corresponds to asymmetrical distributions with longer and heavier tails characteristic of MD distributions. Johnson S_L distributions may be used to estimate any distribution in this region. Also, based on the value of mn/p^2 , the system gives information on distance from the Pearson Type III line or lognormal line. For example, in fitting the Gamma distribution in Figure 3.1, the value is 0.83, indicating that the S_L fit is closer to the Type III line (Gamma distribution locale). In

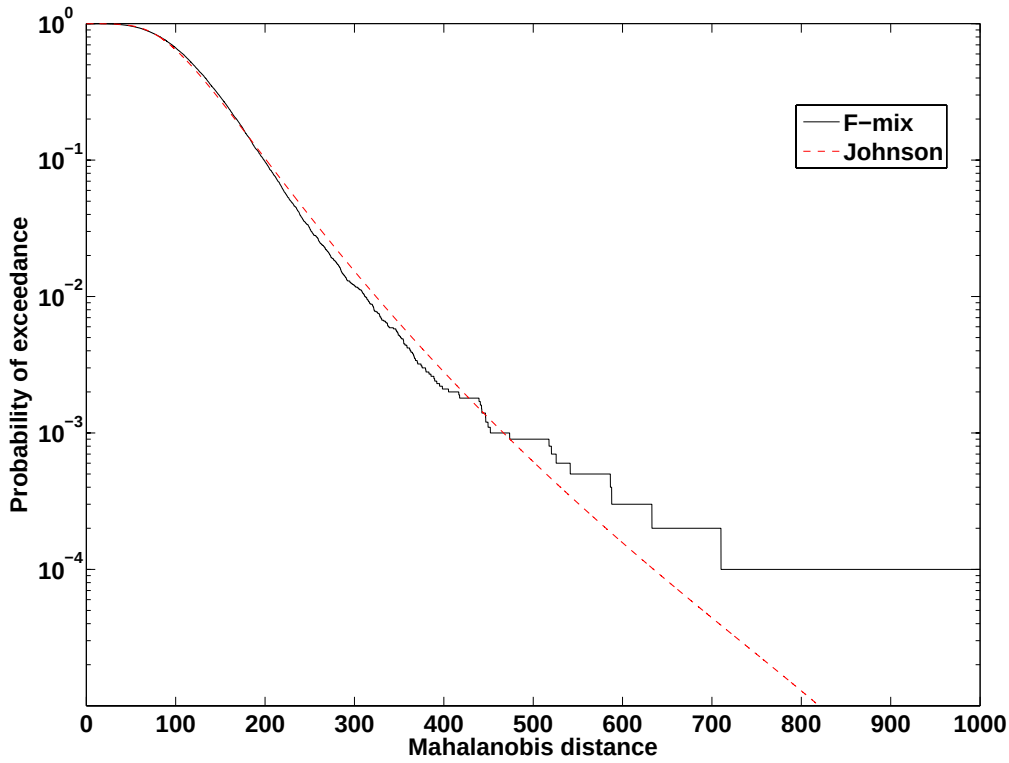


Figure 3.5: Johnson S_L distribution (dotted line) fit to 10,000 values generated from a mixture of two F-distributions with parameters: $\nu_1 = 155$ and $\nu_2 = 100$ for the first F-distribution and $\nu_1 = 155$ and $\nu_2 = 10$ for the second F-distribution, and mixed at 20 % of the first distribution and 80 % of the second (solid line).

fitting the lognormal distribution in Figure 3.3, the value is 1.01, indicating that the S_L fit closely approaches the lognormal line. In Figure 3.5, the value is 1.48, indicating that the mixture of two F-distributions resembles an F-distribution approaching the Type V line. For this feature (which imparts information about the distribution family) and for the same imposed limits, the Johnson S_L system is selected to model HSI MD distributions.

3.2 Fitting HSI MD Distributions with Johnson S_L

Johnson S_L distributions are used to model the MD distributions of five clusters taken from an AVIRIS data collect at Ft A.P. Hill, VA [24]. The image and cluster

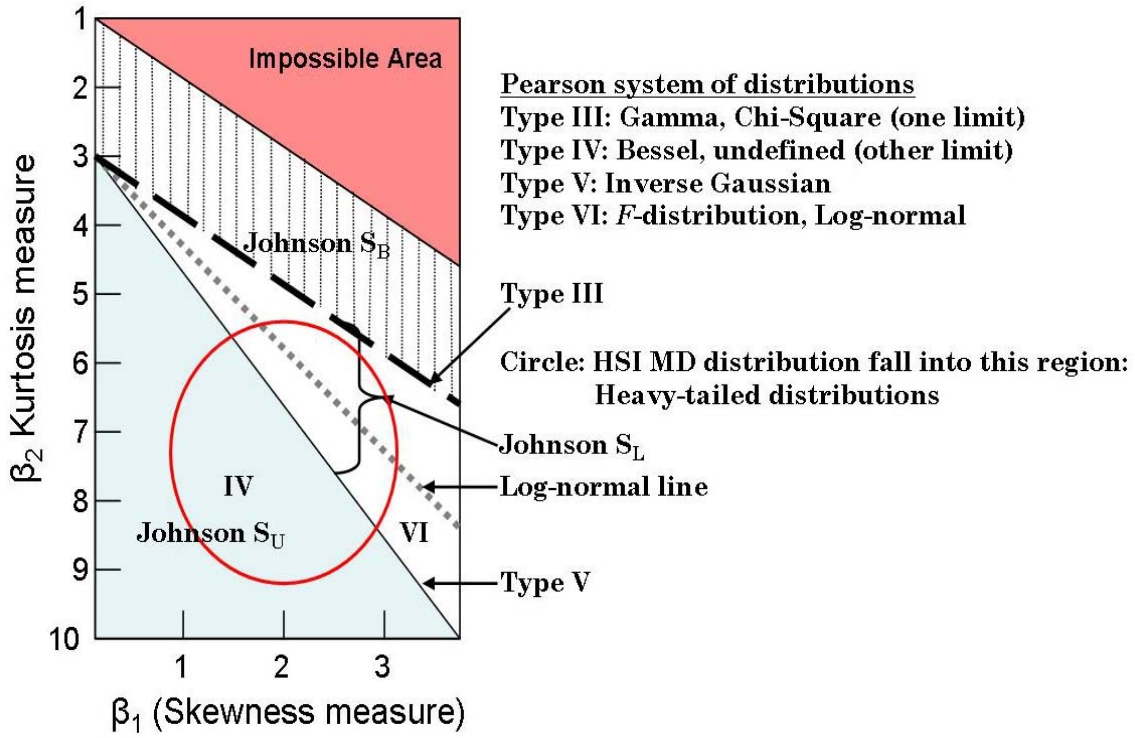


Figure 3.6: A β_1, β_2 chart showing different regions for Johnson system distributions and Pearson system distributions. The lightly shaded lower left triangle represents the Johnson S_U distribution region, The hashed region below the "Impossible Area" shows the Johnson S_B distribution region. The types of distributions fit by the Johnson S_L reside in the region between the Type III line and the Type V line. The circled area represents an area of heavy-tailed distributions where HSI MD distributions tend to be located.

map are shown in Figure 3.7. The clusters are obtained by performing SEM analysis on the data set (see Appendix A), and the image data is clustered into five classes of material. A mixture of two F-distributions is also fit to the data, as in [54], for comparison.

The estimated Johnson S_L distributions are then compared to the F-distribution mixtures using the MSE of the exceedance curves. For example, Figure 3.8 shows the exceedance curve of a mixture of F-distributions and the exceedance curve of a Johnson S_L distribution fit to a MD distribution, where the data is from cluster 1 in Figure 3.7. The MSE between the exceedance curve of the data and the exceedance curve of the mixture of F-distributions is 1.8×10^{-4} , whereas the MSE between the

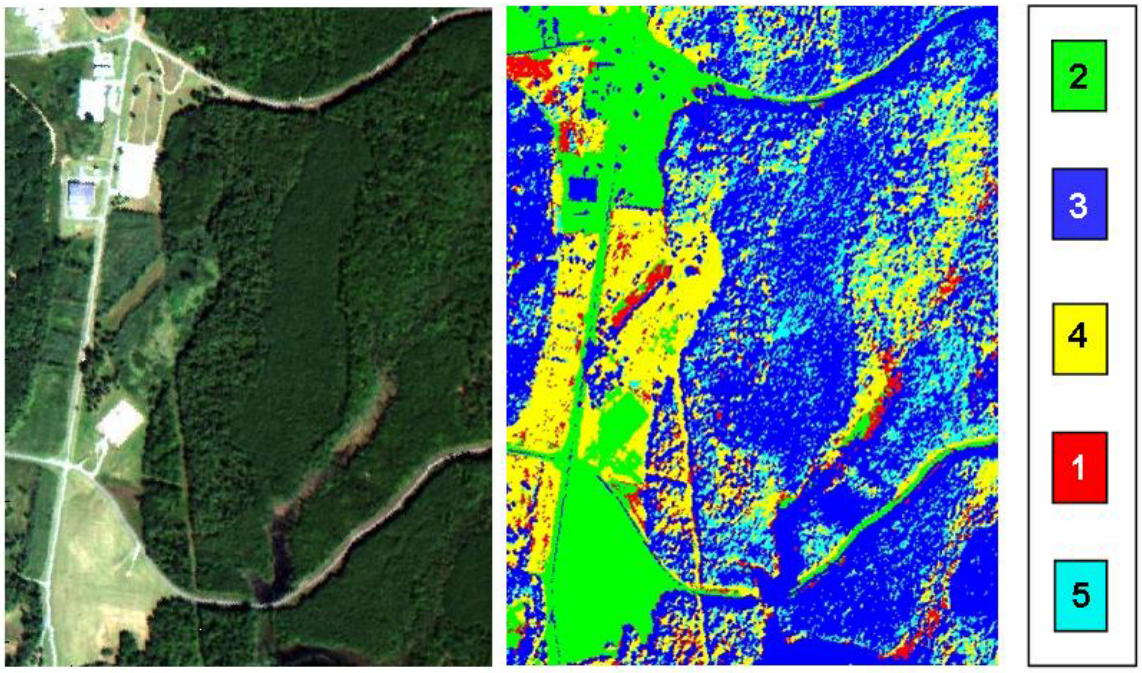


Figure 3.7: Left: False-color image of the Ft A.P. Hill, VA, AVIRIS hyperspectral data. Middle: Image showing pixels grouped into five different clusters by the SEM algorithm. Right: Color key for identifying clusters.

exceedance curve of the data and the exceedance curve of the estimated Johnson S_L distribution is 0.11×10^{-4} .

The exceedance MSE gives a scalar value for a fit to the data set. The plot in Figure 3.8 has a logarithmic scale on the y-axis, and the tail of the fit suggests a large deviation from the Johnson S_L distribution. However, on a linear scale the differences are minute. An exceedance plot of a χ^2 distribution with d degrees of freedom is plotted for comparison to large fitting error (relatively, small fitting errors on this logarithmic scale show a fit closer to the data, while large fitting errors plot closer to the χ^2 exceedance plot).

The comparable accuracy of the F-mixture distribution and the Johnson S_L distribution in Figure 3.8 is evident. However, the Johnson distribution appears to model the body of the data more accurately. The body of the data refers to the higher density region of the distribution (compared to the lower density region in the tails). This better representation of the data in the body, as well as comparable accuracy

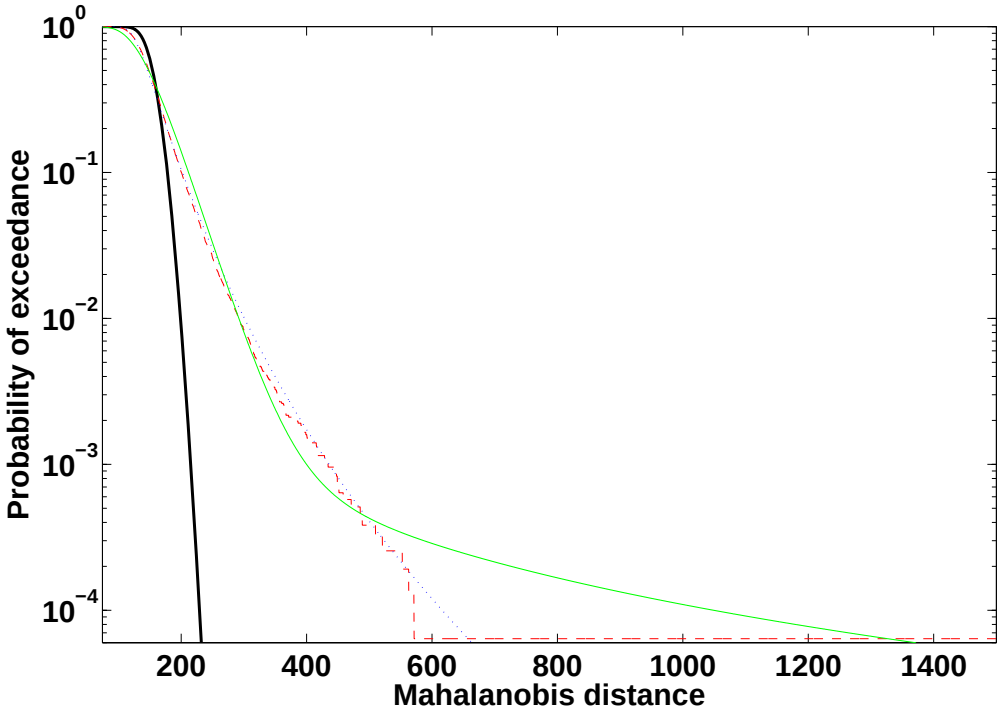


Figure 3.8: Probability of exceedance versus MD for 14,243 data points from cluster 1 (dashed curve), F-mixture distribution (light solid curve), Johnson S_L distribution (dotted curve), and χ^2 distribution (thick curve). Notice that the Johnson S_L distribution performs comparably to the F-mixture distribution.

in the tails, is evident in the more-than-an-order-of-magnitude improvement in the MSE value of the Johnson S_L distribution over the F-mixture distribution exceedance plots. Figure 3.9 shows a zoom-in on the body of the data.

For the remaining four clusters only exceedance plots are given. The y-axis is scaled from one to 10^{-4} on each plot. However, the x-axis varies based on the range of MDs.

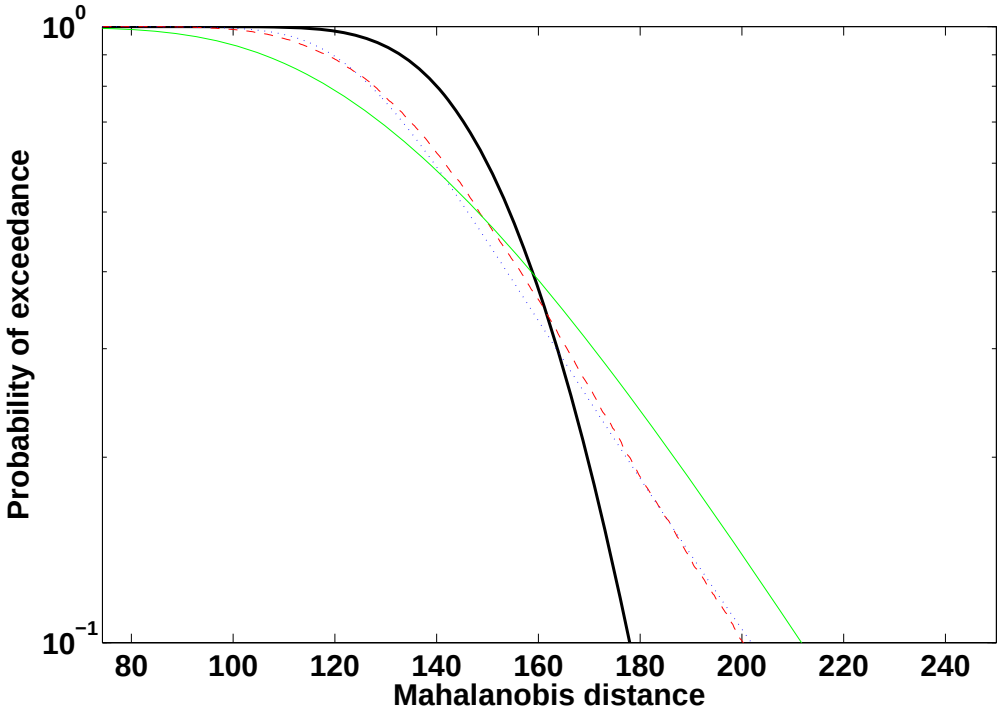


Figure 3.9: The body of the probability of exceedance versus MD for Cluster 1 (dashed curve), F-mixture distribution (light solid curve), Johnson $S_L$distribution (dotted curve), and χ^2 distribution (thick curve). Notice that the Johnson S_L distribution fits closer to the data as expected. The Johnson system parameters are estimated directly from the data, and the MD data gives over 10,000 data points to finely tune the parameter estimates.

3.3 Results from Fitting HSI MDs with Johnson S_L Distributions

From Figures 3.8 - 3.13, note that the Johnson S_L distribution is comparable to the F-mixture distribution for fitting MD distributions. Also, Table 3.1 shows the MSE for each method fit to the data. In the majority of cases the Johnson S_L distribution outperforms the F-mixture distribution by nearly an order of magnitude.

The Johnson S_L distribution performs well in the majority of cases because it fits the body as well as the tail to a high degree of accuracy. The mixture of F-distributions fits the tail well but not the body. However, since the tail of the distribution is where small probability of false alarm thresholds occur, it is advantageous

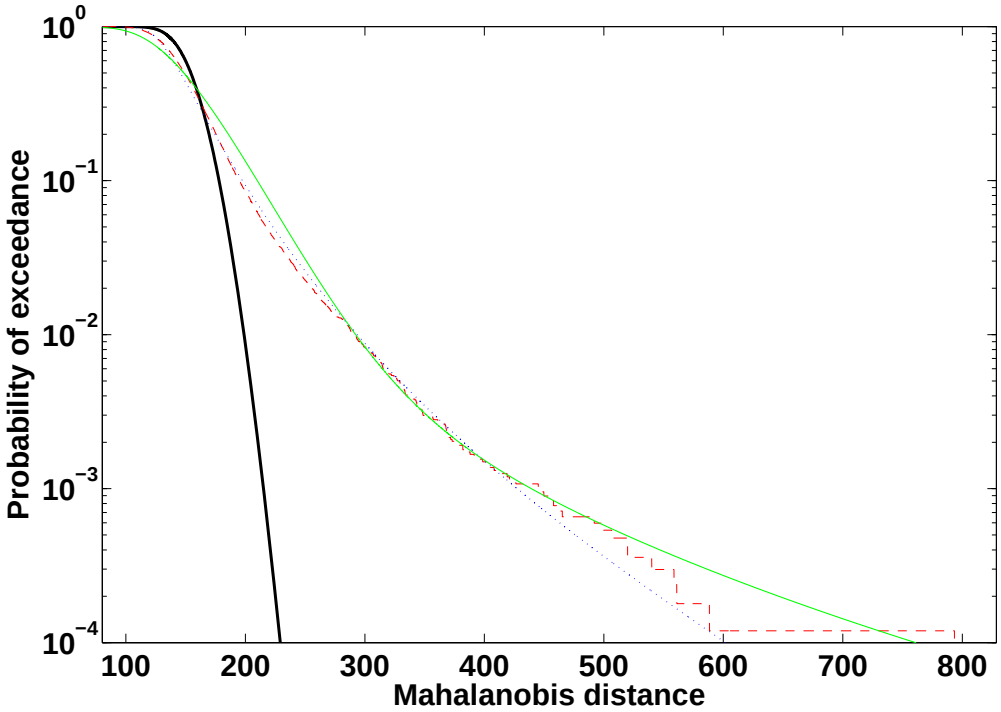


Figure 3.10: Probability of exceedance versus MD for Cluster 2 (dashed curve), F-mixture distribution (light solid curve), Johnson S_L distribution (dotted curve), and χ^2 distribution (thick curve).

Table 3.1: Summary of MSE for Johnson S_L Distribution and F-distribution Mixture Fit to MD Data.

Cluster Number	Mean Square Error F-distribution Mixture ($\times 10^{-4}$)	Mean Square Error Johnson Distribution ($\times 10^{-4}$)
1	5.14	1.11
2	1.62	0.25
3	6.88	1.01
4	5.85	1.11
5	2.77	0.62

to modify the fitting metric such that the fit at the tails is weighted more than the body. Therefore the MSE is modified as

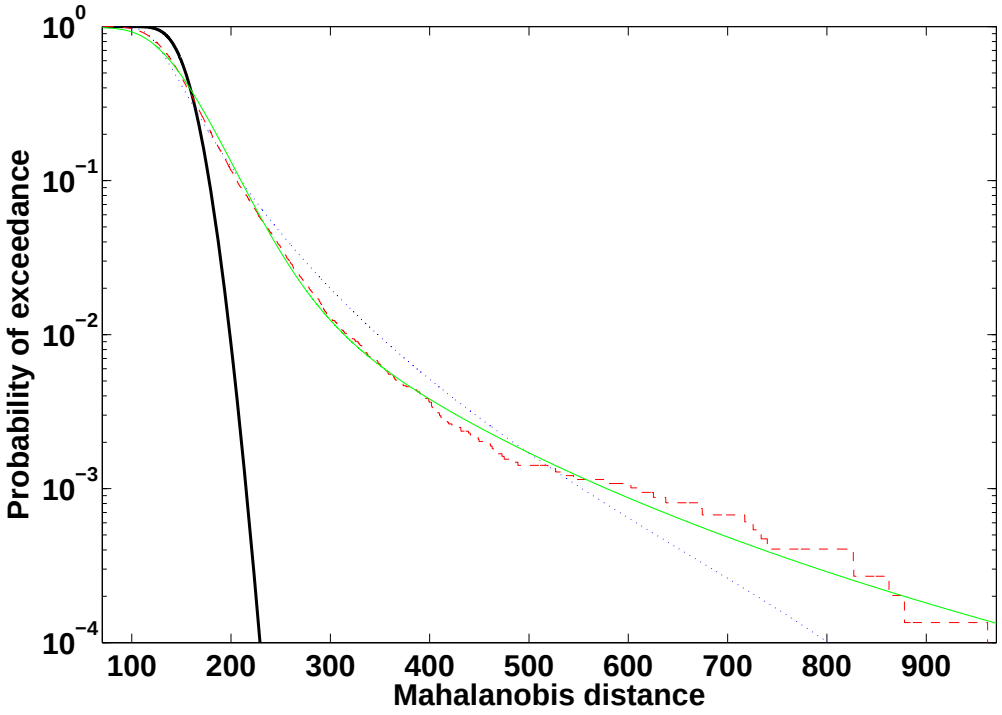


Figure 3.11: Probability of exceedance versus MD for Cluster 3 (dashed curve), F-mixture distribution (light solid curve), Johnson S_L distribution (dotted curve), and χ^2 distribution (thick curve).

$$MSE = \frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2\,(x_i)^{\beta_1}, \tag{3.4}$$

where x_i are the MD values, y_i are the data exceedance values at x_i, $\hat{y}_i$ are the fitting function exceedance values at x_i, and β_1 is the skewness value for the data. Notice from Figure 3.6 that the region where HSI MD data distributions occur result in β_1 values on the order of 1 to 3, with larger skewness values implying heavier tails. Therefore, the MSE metric is modified adaptively, according to the thickness of the tail of the MD distribution. Figure 3.14 shows two plots of β_1 values common to heavy-tailed distributions. Notice that the larger β_1 value results in higher weights for MD values in the extremities of the tail. The result is that weights on the MSE around the body region are smaller than weights on MSE values at greater extremities

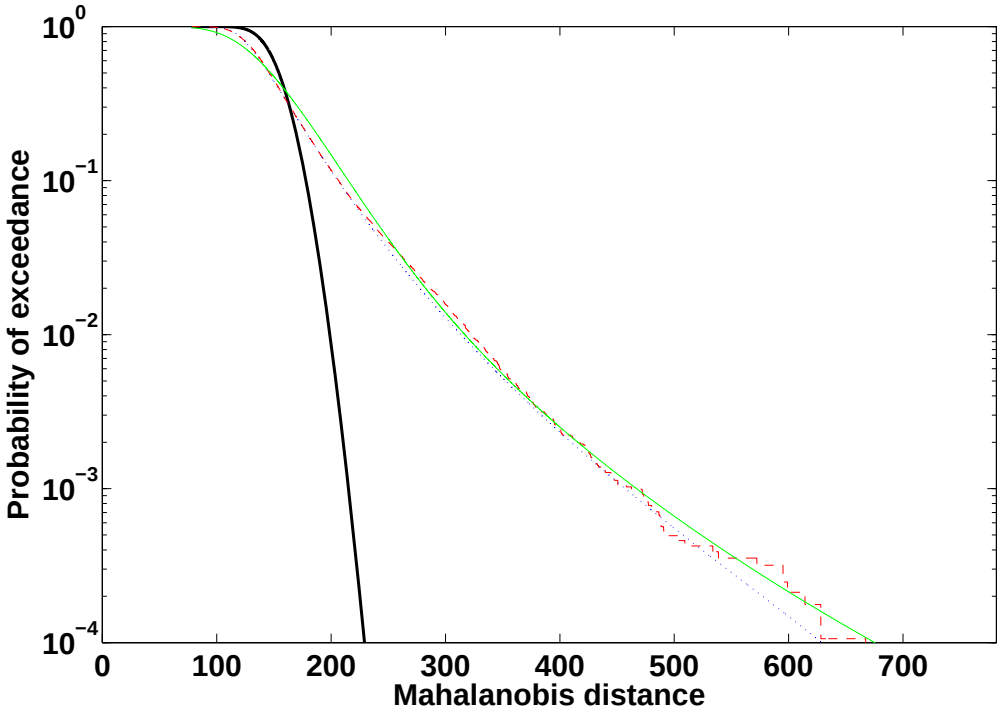


Figure 3.12: Probability of exceedance versus MD for Cluster 4 (dashed curve), F-mixture distribution (light solid curve), Johnson S_L distribution (dotted curve), and χ^2 distribution (thick curve).

of the tail. Specifically, the weighting term $(\cdot)^{\beta_1}$ compensates for the fact that heavy-tailed distributions decrease geometrically [3, 36]

Table 3.2 shows the performance of each distribution fit to the data using the weighted MSE metric. Notice that the new metric for fitting yields the same conclusion; for the majority of the fits the Johnson S_L distribution fit outperforms the F-mixture fit.

Also of importance in comparing the two modeling techniques, and one of the evaluation criteria in achieving the objective for this part of the research, is computational efficiency. Whereas the determination of F-mixture distributions requires minutes of computational time on a conventional Pentium 4 processor, the Johnson S_L distribution is derived within seconds of data input. This increase in computational efficiency is important for determining the best algorithm for data analysis.

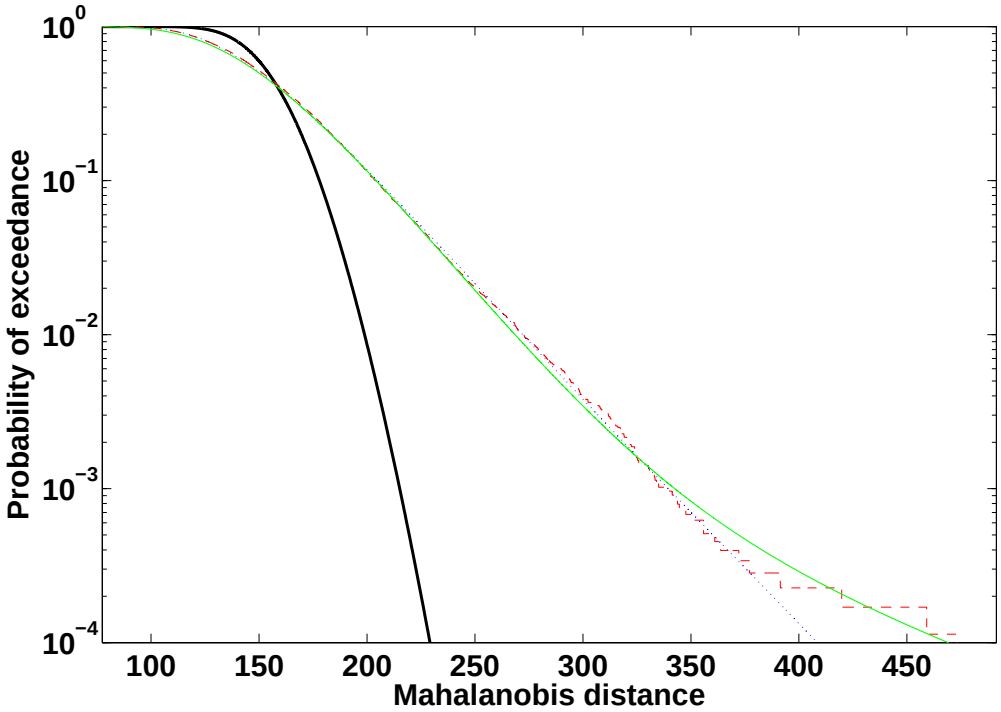


Figure 3.13: Probability of exceedance versus MD for Cluster 5 (dashed curve), F-mixture distribution (light solid curve), Johnson S_L distribution (dotted curve), and χ^2 distribution (thick curve).

Table 3.2: Summary of tail-weighted MSE for Johnson S_L Distribution and F-distribution Mixture Fit to MD Data.

Cluster Number	Tail-weighted Mean Square Error F-distribution Mixture ($\times 10^{-3}$)	Tail-weighted Mean Square Error Johnson Distribution ($\times 10^{-3}$)
1	1.03	0.87
2	1.12	0.99
3	0.85	0.86
4	0.66	0.62
5	0.94	0.79

HSI data exploitation demands the manipulation of large data sets over short periods of time, and computationally efficient routines are more applicable.

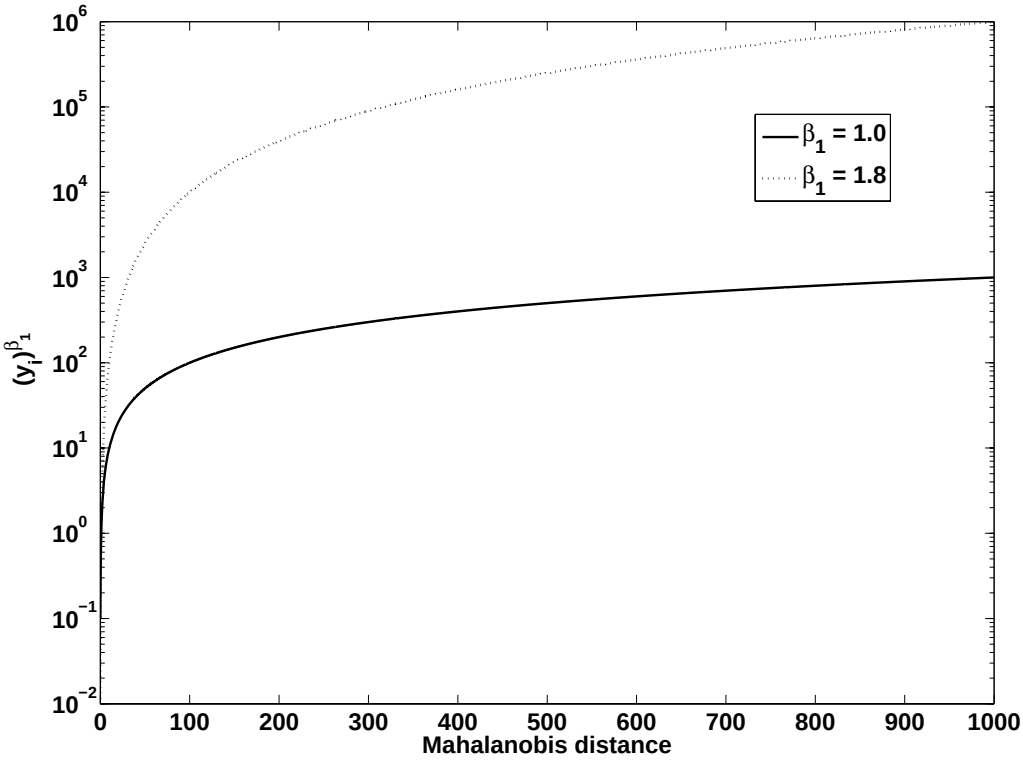


Figure 3.14: Weighting on MD values for the modified MSE metric. The modified MSE is weighted such that tail values (especially extreme tail values) in the distribution fit error are penalized more than errors in the fit to the body

3.4 Improving the Johnson S_L Distribution Fit for Robustness Against Perturbations

3.4.1 Definition of a Perturbation in the Data. Currently, the effect of perturbations is not taken into account, which can be detrimental to the performance of the Johnson S_L distribution fit. Here, perturbations in the decrease of an exceedance function are termed "outliers," i.e., anomalous observations that show up as "outlying" data points with respect to the shape of the exceedance function. In a standard statistical definition for univariate data, outliers are points that lie outside of three standard deviations from the mean of the distribution [60]. However, this definition ignores data naturally located in the extremities of the tails of the distribution.

In most cases, an adaptive method is needed which takes into account sample size, non-Gaussian nature of the data, and robust measures of mean and covariance.

Here these factors are taken into account by examining outlier effects on the exceedance curves of the MD distributions. Generally, outliers are observations resulting from a secondary process and not extreme values ensuing from the cluster distribution. These secondary processes are seen as multi-modality in what should be unimodal MD distributions [54]. Multi-modality is noticeable in exceedance plots as a change in the rate of decrease of the exceedance curve of the data.

Thus outliers suggest multi-modality in the exceedance curve of cluster MD distributions. This multi-modality is evidenced as a change in first and second derivatives of a curve. The first derivative decreases as the rate of descent changes direction. However, for a rate of descent common to most distribution tail regions, the second derivative of the curve is positive (i.e., the curve is concave up). When the second derivative of the curve decreases and then increases again, the exceedance curve has an unnatural "bump" caused by an outlier. A type of "bump" in the exceedance curve is shown in Figure 3.15.

Taking into account this phenomenon, an approach to determining outliers for HSI clusters called "Leave-One-Out-Smoothness" (LOOS) is used here. LOOS is based on using the integrated squared second derivative to measure the smoothness of an exceedance curve. Random data points from a specific region in the MD distribution are left out (hence, "Leave-One-Out-Smoothness") in order to find the smoothest representation. Once this representation is determined, data points left in the cluster are analyzed for Minimum Covariance Determinant (MCD).

The MCD algorithm estimates the proper size of data samples in a cluster with respect to leaving out possible outliers [68, 69]; it determines the proper subset of data of size g that minimizes the determinant of the sample covariance matrix [22]. Size g is usually specified by the user (based on previous analysis) with the correct

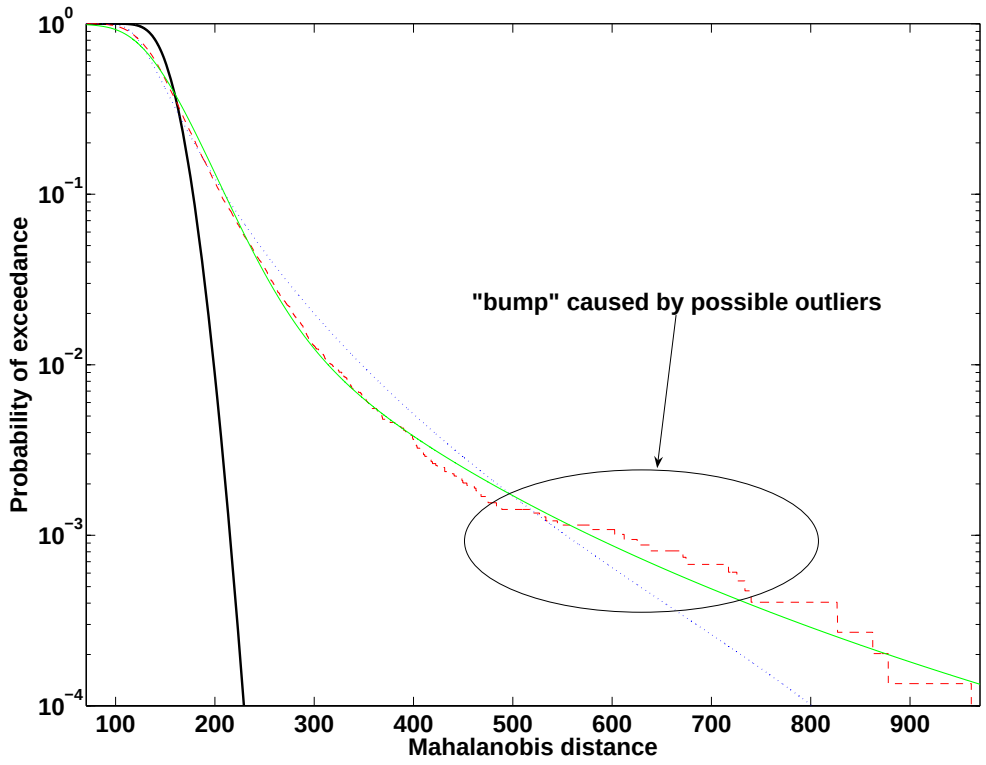


Figure 3.15: An example of exceedance curve behavior under the influence of "outliers".

data members of size g determined by MCD to produce a robust model. Here g is determined by the number of data points left by LOOS.

3.4.2 Proper "Outlier" Regions. For computational efficiency, the entire exceedance curve need not be examined in implementing LOOS. Out of the approximately 20,000 data points which normally comprise a MD distribution exceedance curve, only a portion of the tail needs analysis. Here the portion examined is near the 10^{-3} region on the exceedance curve; outliers are examined only in this area due to their significant impact on the shape of the distribution. For example, a subset of MDs from cluster 4 in Figure 3.7 is distorted by the addition of random outliers in three regions of the tail: exceedance values at $10^{-2} \pm 0.005$, $10^{-3} \pm 0.0005$, and $10^{-4} \pm 0.00005$. Varying numbers of outliers are placed in these regions, and the be-

havior of Johnson distributions in fitting the exceedance curve is tracked. The changes in the parameters associated with the Johnson distributions are given in Figures 3.16, 3.17, and 3.18. Only the behaviors of the η_J and γ_J values are tracked in this analysis because these values are associated with the shape of the distribution (the ϵ_J and λ_J values are linear translation variables that do not alter shape).

Also, in the figures S_U and S_L Johnson distributions are used to monitor parameter change to emphasize the fact that a distribution will change its "type" (Pearson type/Johnson type) when secondary process data infiltrate the distribution of the process of interest. In practice one can use the Johnson S_U distribution to model MD data by shifting the distribution so that its left tail returns values very close to zero beyond x-axis ≤ 0. However, in principle only the Johnson S_L distribution should be used to model MD data.

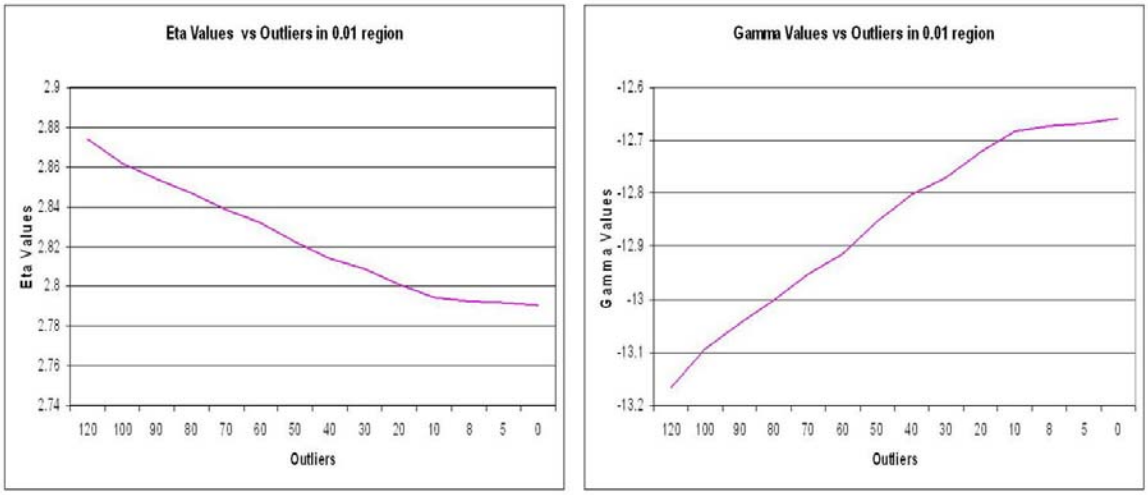


Figure 3.16: The η_J and γ_J values of Johnson distributions fit to an MD distribution (see text) with a varying number of outliers in the $10^{-2} \pm 0.005$ region of exceedance.

From the plots it is evident that outliers in the 10^{-2} region do not greatly affect the shape of a distribution. The small changes in η and γ do not significantly alter the shape of the Johnson distribution type. But outliers in the 10^{-4} region have a great impact. However, outliers in this region and beyond are "legitimate outliers," in that they reveal information about a secondary process of interest in the cluster. For example, a target partially hidden in foliage in an HSI scene generates an extreme

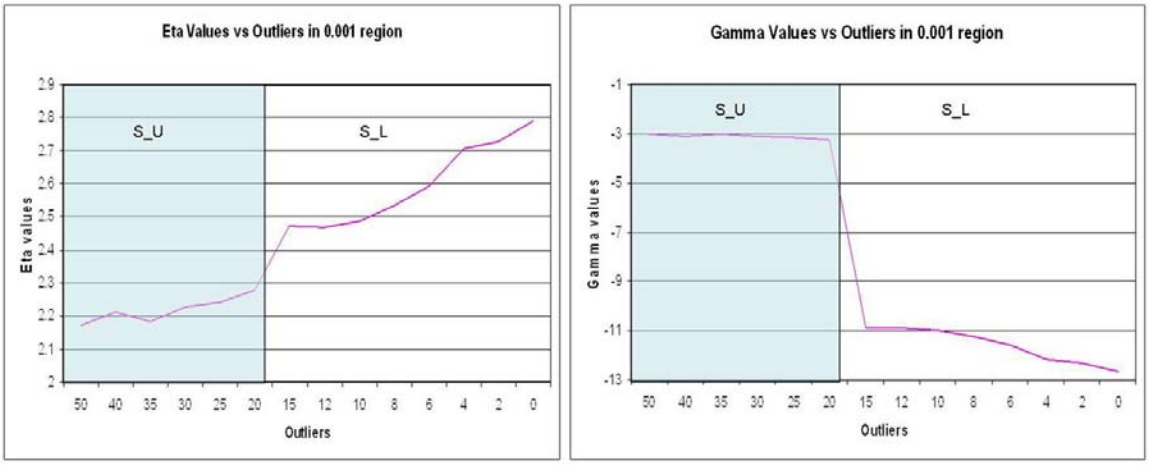


Figure 3.17: The η_J and γ_J values of Johnson distributions fit to an MD distribution (see text) with a varying number of outliers in the $10^{-3} \pm 0.0005$ region of exceedance.

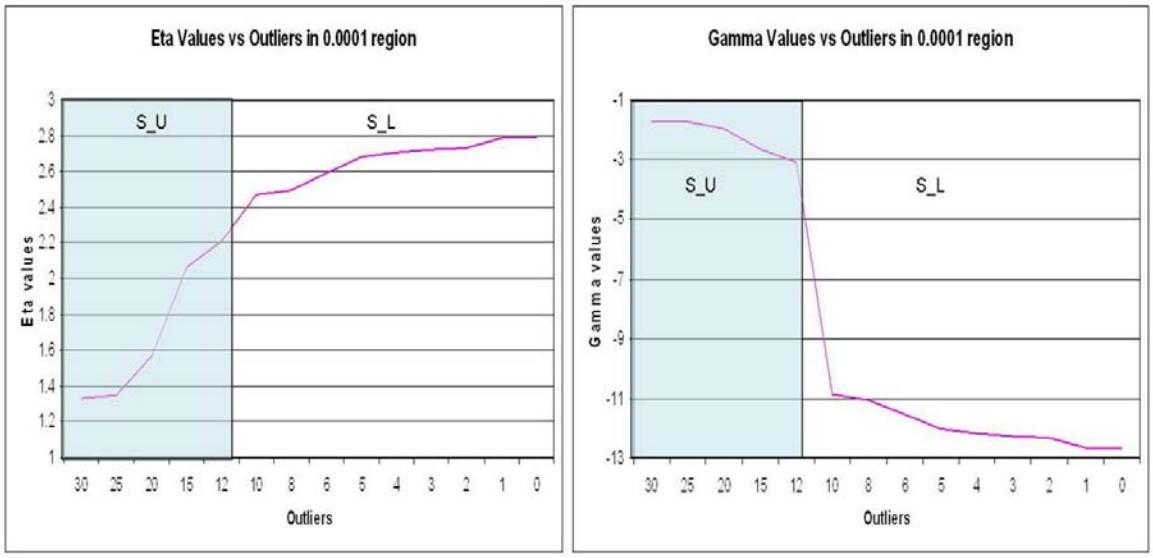
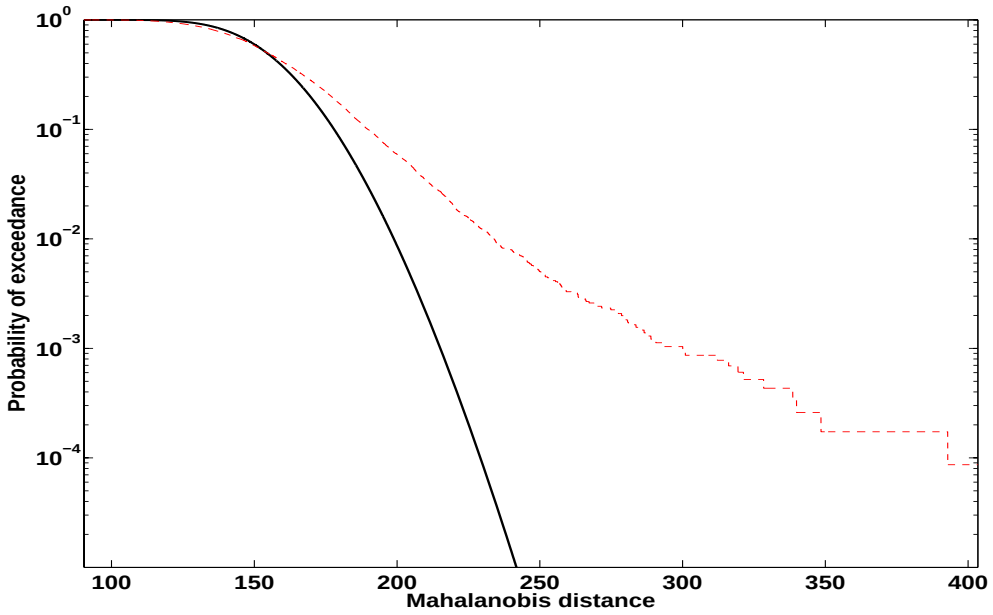
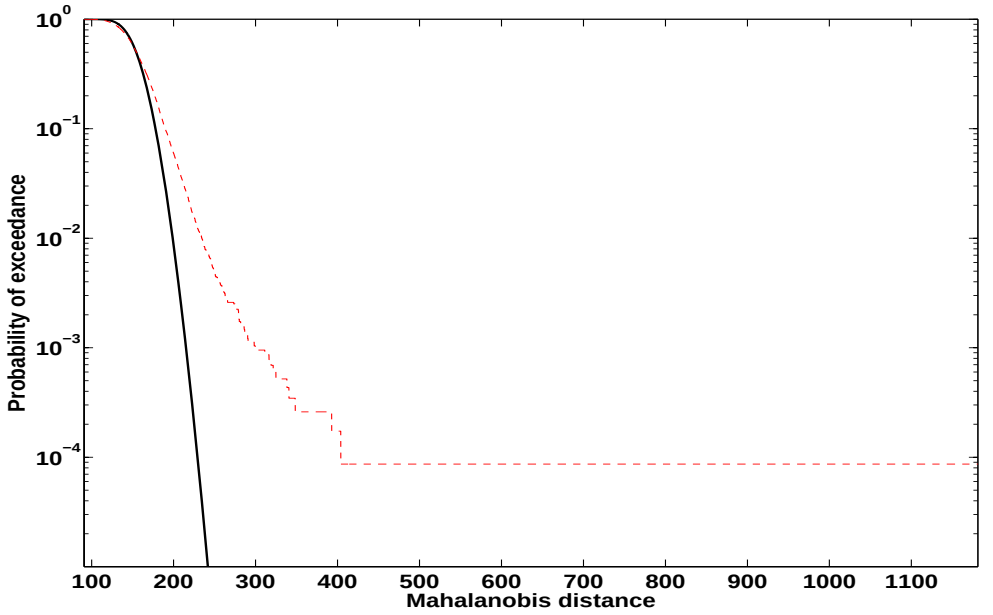


Figure 3.18: The η_J and γ_J values of Johnson distributions fit to an MD distribution (see text) with a varying number of outliers in the $10^{-4} \pm 0.00005$ region ofexceedance.

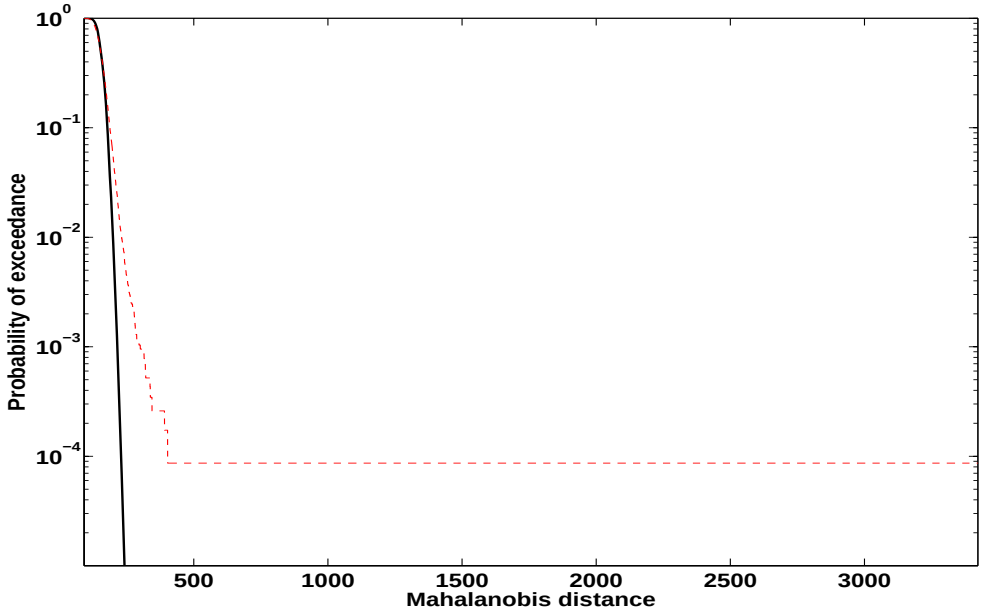
outlier data point in the MD distribution (exceedance of 10^{-4} and beyond), as is shown in the four exceedance curves in Figure 3.19, where a synthetic pixel is inserted into a vegetative cluster (a subset of cluster 3). The synthetic pixel is a metallic spectrum from an airplane in Figure 2.18 gradually mixed with tree spectra (simulating partial coverage under foliage at different levels).


Probability of exceedance
Mahalanobis distance

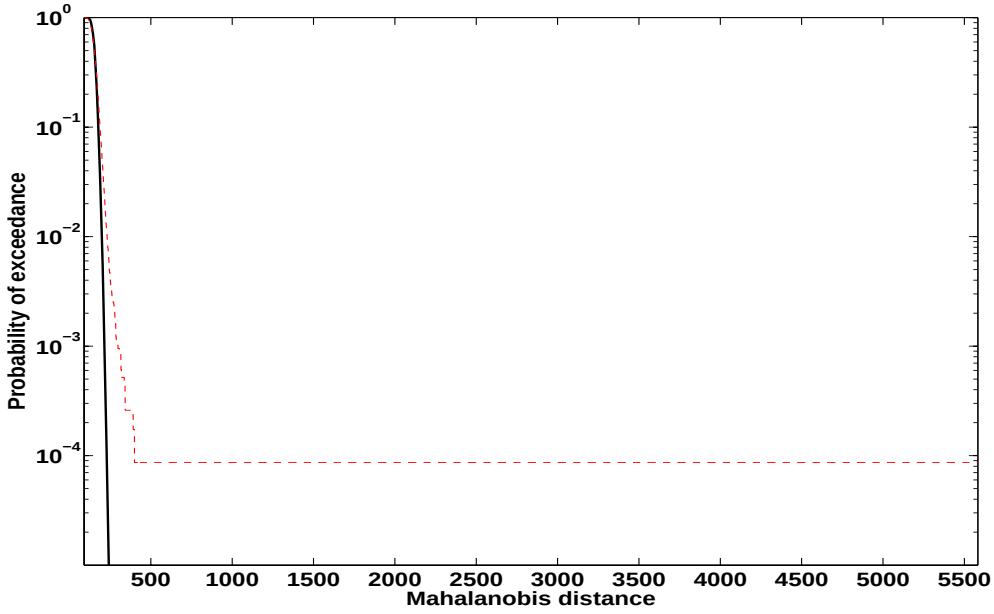
(a)


Probability of exceedance
Mahalanobis distance

(b)



(c)



(d)

Figure 3.19: Target spectrum mixed with a background spectrum and inserted into a cluster of tree spectra (see text) at (a) 0 % target, 100 % background, (b) 20 % target, 80 % background, (c) 40 % target, 60 % background, and (d) 60 % target, 40 % background. The dashed line represents the exceedance curve of the data and the solid line is a χ^2 exceedance curve.

Cases (c) and (d) in Figure 3.19 are extreme cases but can occur due to SEM redistribution of pixel spectra from clusters deemed too small for proper parameter estimation. However, the more likely case (b) shows the difference one small alteration in a pixel spectrum can cause with respect to the background MD distribution. If five or more of these types of altered pixels appear in the cluster, then these data cause a significant change in the shape of the tail extremity. However, as mentioned previously, these are the "outliers" of interest.

"Outliers" at $10^{-3} \pm 0.0005$ exceedance have a noticeable impact on the shape of the distribution. As indicated in the figures, perturbations in this region can alter the η_J and γ_J parameters so that the type of distribution changes from S_L to S_U (two different families of distributions). Therefore, LOOS is applied to outliers in this region.

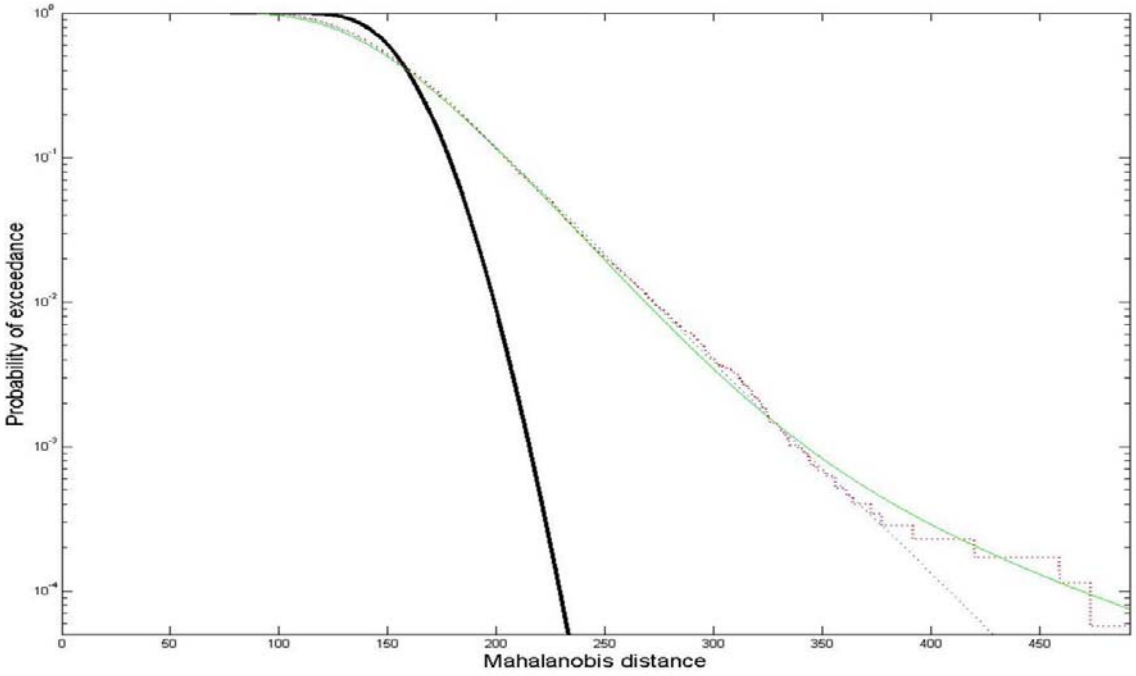


Figure 3.20: Original exceedance plot of subset of MDs from cluster 4 in Figure 3.7. The Johnson distribution fitting this data is an S_L distribution with parameters γ_J = -12.66 and $\eta_J = 2.79$. The mixture of F-distributions fitting this data has parameters $\nu_1 = 50$, $\nu_2 = 16$, and $w = 0.97$. The MCD for the cluster data is 1.14.

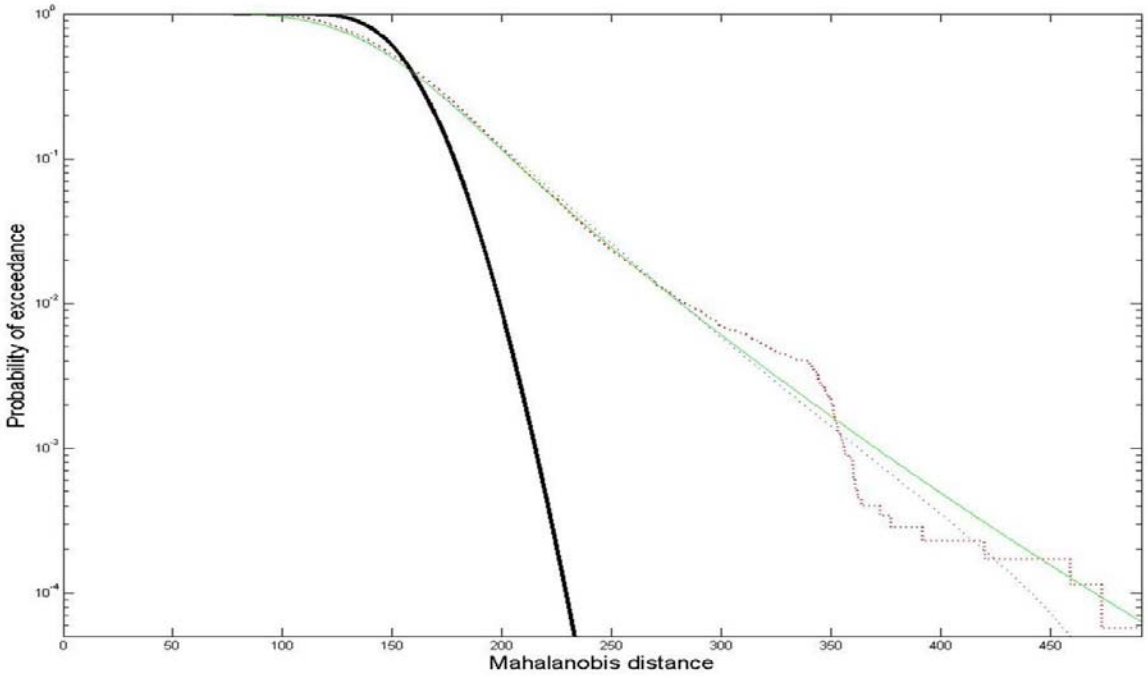


Figure 3.21: Fifty simulated outliers added at $10^{-3} \pm 0.0005$ exceedance. The Johnson distribution fitting this data is S_U with parameters γ_J = -3.10 and η_J = 2.21. The mixture of F-distributions fitting this data has $\nu_1 = 30$, $\nu_2 = 100$, and w = 0.56. Cluster MCD is 40.04.

3.4.3 Mitigating "Outlier" Data. The MCD for the subset of MDs from cluster 4 in Figure 3.7 is calculated to verify that the smoothing creates a decrease in the determinant of the covariance for the cluster of data. For the example presented here, simulated outliers are added to the exceedance plot of the MD distribution from a cluster of hyperspectral data. Then LOOS is applied to exceedance data and the MCD is calculated for pre-smoothed and smoothed cluster data. The results are shown in Figures 3.20, 3.21, and 3.22.

By applying the LOOS smoothing to the exceedance curve affected by the outliers, distortion in the shape of the Johnson and F-mixture distributions is mitigated. Note that insertion of outliers greatly changes the parameter values for both the Johnson and F-mixture distributions (the Johnson distribution type also changes from S_L to S_U). Note that in eliminating some of the outliers (as a result of applying LOOS), the parameters are driven back toward their original values.

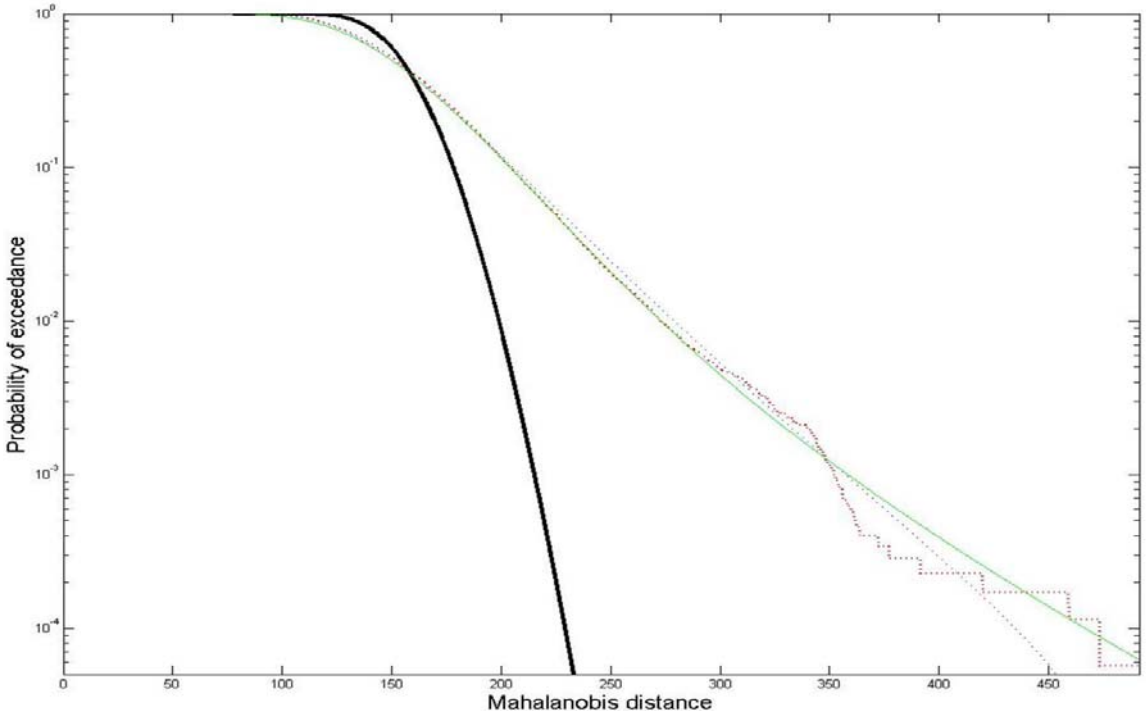


Figure 3.22: Exceedance plot from Figure 3.21 smoothed by LOOS. The Johnson distribution fitting this data is an S_L distribution with parameters γ_J = -12.83 and $\eta_J = 2.47$. The mixture of F-distributions fitting this data has $\nu_1 = 60$, $\nu_2 = 26$, and $w = 0.77$. Cluster MCD is 10.50.

3.5 Multivariate EC Model from the Univariate Johnson S_L

Objective 2 is developed for determining the multivariate model of the HSI data given the univariate MD information. Here the theory of EC distributions (sometimes referred to as multivariate symmetric distributions [21]) is used to obtain a multivariate HSI data model from the univariate Johnson S_L distribution fit to the MD data. Once the multivariate data model is determined, d-dimensional data based on this model is generated.

A multivariate EC distribution is characterized by a d-dimensional probability density function (PDF)

$$p(x|\Sigma) = c\,|\Sigma|^{-\frac{1}{2}}\, g(\Delta^2), \tag{3.5}$$

where Σ is the covariance matrix, c is a normalizing constant, Δ is MD, and g is a specified function. The normal distribution is a special case of this family for which $g(\Delta^2) = e^{-\frac{\Delta^2}{2}}$.

The distribution of MDs may be used to determine the distribution of a multivariate data set. For example, MDs are distributed as univariate Chi-squared with d degrees of freedom for a population of d-dimensional multivariate normal random vectors [34]. Thus the normality of multivariate data can be tested by observing the behavior of its MD distribution.

Manolakis *et al.* [54], show that the distribution of MDs obtained from clusters in HSI data follow a mixture of univariate F-distributions. This result implies that the multivariate data follow a mixture of elliptical t-distributions (another member of the EC distribution family). The mixture of F-distributions is developed by weighting an F-distribution that models the body of the MDs with another F-distribution that models the behavior of the tails.

3.5.1 EC Distribution Theory. Fang, et al. [21] state that a $d \times 1$ random vector $\mathbf{x}$ has an EC distribution (contours of equal probability are concentric ellipses about a mean value) with parameters $\boldsymbol{\mu}(d \times 1)$ and $\boldsymbol{\Sigma}(d \times d)$ (mean vector and covariance matrix, respectively) if

$$\mathbf{x} \sim \boldsymbol{\mu} + \mathbf{A}^T\mathbf{y}, \qquad \mathbf{y} \sim S_d(\phi), \tag{3.6}$$

where $\mathbf{A}$ is a $(n \times d)$ matrix such that $\mathbf{A}^T\mathbf{A} = \boldsymbol{\Sigma}$, and $S_d(\phi)$ denotes a d-dimensional spherically symmetric distribution (spherical distribution) with characteristic generator ϕ. The characteristic generator is a scalar function such that

$$\psi(\mathbf{t}) = \phi(\mathbf{t}'\mathbf{t}), \tag{3.7}$$

where $\psi(t)$ is the characteristic function of a spherical distribution. A spherical distribution is defined such that for a $d \times 1$ random vector $\mathbf{y}$,

$$\Omega \mathbf{y} \sim \mathbf{y} \tag{3.8}$$

for every Ω, where Ω is a $d \times d$ orthonormal operator. Similarly, $\mathbf{y}$ is spherically distributed if

$$\mathbf{y} \sim r\mathbf{U}^d, \tag{3.9}$$

where r is a random variable indicating distance (radius) in spherical coordinates, independent of the d-dimensional random unit vector $\mathbf{U}^d$, which is uniformly distributed on the unit sphere in $\Re^d$.

As an example from [21] of an EC distribution, if $\mathbf{y} \sim N_d(\mathbf{0}, \mathbf{I}_d)$, where $N_d(\mathbf{0}, \mathbf{I}_d)$ denotes a d-dimensional normal distribution, or, equivalently, $\mathbf{y} \sim S_d(\phi)$ with $\phi(u) = \exp(-\frac{u}{2})$, $\boldsymbol{\mu}$ is $d \times 1$, and $\mathbf{A}$ is $d \times d$ such that $\mathbf{A}^T\mathbf{A} = \Sigma$, then

$$\mathbf{x} \sim \boldsymbol{\mu} + \mathbf{A}^T\mathbf{y} \tag{3.10}$$

has an EC distribution $\mathbf{x} \sim EC_d(\boldsymbol{\mu}, \Sigma, \phi)$. For this example $\mathbf{x}$ is multivariate normal with $\Delta \sim \chi_d$ (χ-distribution with d degrees of freedom). Likewise, Δ^2 has a χ^2-distribution.

From [21] and [8], if $\mathbf{x} \sim r\mathbf{U}^d$, then $\mathbf{x}$ has a density generator $g(\Delta^2)$, which defines a density for $cg(\Delta^2)$, where c is a normalizing constant, such that Δ^2 has a density

$$f(\Delta^2) = \frac{\pi^{\frac{d}{2}}}{\Gamma(\frac{d}{2})}(\Delta^2)^{d/2-1}g(\Delta^2), \tag{3.11}$$

where $\Gamma(\cdot)$ is the Gamma function. For densities of random vectors $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n$,

$$\Delta^2 = (\mathbf{x} - \boldsymbol{\mu})^T \Sigma^{-1} (\mathbf{x} - \boldsymbol{\mu}). \tag{3.12}$$

One can then write $\mathbf{x} \sim EC_d(\boldsymbol{\mu}, \Sigma, g)$, where $\mathbf{x} \sim \boldsymbol{\mu} + \mathbf{A}^T\mathbf{y}$ and

$$\mathbf{x} \sim C_d |\Sigma|^{-\frac{1}{2}} g((\mathbf{x} - \boldsymbol{\mu})^T \Sigma^{-1} (\mathbf{x} - \boldsymbol{\mu})), \tag{3.13}$$

where C_d is the normalizing factor [21]

$$C_d = \frac{\Gamma(\frac{d}{2})}{\pi^{\frac{d}{2}}} \int_0^{+\infty} (\Delta^2)^{d/2-1} g(\Delta^2) d\Delta^2. \tag{3.14}$$

If $C_d = (2\pi)^{-\frac{d}{2}}$, $g(\Delta^2) = \exp(-\frac{\Delta^2}{2})$, $\Sigma = \mathbf{I}_d$, and $\boldsymbol{\mu} = 0$, then $\mathbf{x}$ is a multivariate normal distribution.

Note that $(\mathbf{x} - \boldsymbol{\mu})^T \Sigma^{-1} (\mathbf{x} - \boldsymbol{\mu})$ is the term for the MD measure. Therefore, one can use Equation (3.11) to determine the configuration of the density generator $g(\Delta^2)$ from the univariate distribution of MDs, and, in turn, an expression for the multivariate distribution using Equation (3.13).

For example, consider random vectors with MDs following an F-distribution

$$MD = \Delta^2 = (\mathbf{x} - \boldsymbol{\mu})^T \Sigma^{-1} (\mathbf{x} - \boldsymbol{\mu}) \sim F_{d,\nu}(\frac{\Delta^2}{d} \frac{\nu}{\nu - 2}), \tag{3.15}$$

where

$$F_{d,\nu}(\frac{\Delta^2}{d} \frac{\nu}{\nu - 2}) \sim \frac{\Gamma(\frac{d+\nu}{2})}{\Gamma(\frac{d}{2})\Gamma(\frac{\nu}{2})} \nu^{\frac{-d}{2}} (\Delta^2)^{\frac{d}{2}-1} (1 + \frac{\Delta^2}{\nu})^{-\frac{d+\nu}{2}}. \tag{3.16}$$

Using 3.11 yields

$$\frac{\Gamma(\frac{d+\nu}{2})}{\Gamma(\frac{d}{2})\Gamma(\frac{\nu}{2})} \nu^{\frac{-d}{2}} (\Delta^2)^{\frac{d}{2}-1} (1 + \frac{\Delta^2}{\nu})^{-\frac{d+\nu}{2}} = \frac{\pi^{\frac{d}{2}}}{\Gamma(\frac{d}{2})} (\Delta^2)^{\frac{d}{2}-1} g(\Delta^2). \tag{3.17}$$

Substituting $z = \Delta^2$

$$cg(z) = \frac{\Gamma(\frac{d+\nu}{2})}{\Gamma(\frac{\nu}{2})}(\pi\nu)^{\frac{-d}{2}}(1+\frac{z}{\nu})^{-\frac{d+\nu}{2}}. \tag{3.18}$$

The normalizing constant C_d is

$$C_d = \frac{\Gamma(\frac{d}{2})}{\pi^{\frac{d}{2}}}\left[\int_0^{+\infty}(z)^{\frac{d}{2}-1}(1+\frac{z}{\nu})^{-\frac{d+\nu}{2}}dz\right]^{-1}. \tag{3.19}$$

With the substitution $u = 1 + \frac{z}{\nu}$ in the integration term, $du = \nu^{-1}dz$, and the expression in the brackets becomes

$$\nu^{\frac{d}{2}}\int_0^{+\infty}\left(1-u^{-1}\right)^{\frac{d}{2}-1}u^{\frac{-\nu}{2}-1}du.$$

Letting $p = (1-u^{-1})$ results in

$$\nu^{\frac{d}{2}}\int_0^{+\infty}p^{\frac{d}{2}-1}\left(\frac{1}{1-p}\right)^{\nu+2}dp,$$

which results in [23]

$$\nu^{\frac{d}{2}}\frac{\Gamma(\frac{d}{2}))\Gamma(\frac{\nu}{2})}{\Gamma(\frac{d+\nu}{2})}.$$

Therefore,

$$\begin{aligned} C_d &= \frac{\Gamma(\frac{d}{2})}{\pi^{\frac{d}{2}}}\left[\nu^{\frac{d}{2}}\frac{\Gamma(\frac{d}{2}))\Gamma(\frac{\nu}{2})}{\Gamma(\frac{d+\nu}{2})}\right]^{-1} \\ &= \frac{\Gamma(\frac{d+\nu}{2})}{(\pi\nu)^{\frac{d}{2}}\Gamma(\frac{\nu}{2})}, \end{aligned}$$

and thus

$$g(\Delta^2) = \left(1 + \frac{\Delta^2}{\nu}\right)^{-\frac{d+\nu}{2}}, \tag{3.20}$$

which is a common representation of the generating density for a multivariate t-distribution [2]

$$\mathbf{x}_t \sim \frac{\Gamma(\frac{d+\nu}{2})}{(\pi\nu)^{\frac{d}{2}}\Gamma(\frac{\nu}{2})}|\Sigma|^{-\frac{1}{2}}\left(1 + \frac{\Delta^2}{\nu}\right)^{-\frac{d+\nu}{2}}. \tag{3.21}$$

3.5.2 Multivariate EC Density from Univariate Johnson S_L Density. Following the same steps, a multivariate EC density function may be derived as a result of a MD distribution described by a Johnson S_L distribution. For example, given the Johnson S_L density function

$$f(\Delta^2)_{S_L} = \frac{\eta_J}{\sqrt{2\pi}(\Delta^2 - \epsilon)}\exp\left(-\frac{1}{2}\eta_J^2\left[\frac{\gamma_*}{\eta_J} + ln(\Delta^2 - \epsilon_J)\right]^2\right), \tag{3.22}$$

where $\gamma_* = \gamma_J - \eta_J\ ln(\lambda_J)$ from Equation (3.2), and using Equation (3.11), the form of $g(\Delta^2)$ is

$$\frac{\eta_J}{\sqrt{2\pi}(\Delta^2 - \epsilon_J)}\exp\left(-\frac{1}{2}\eta_J^2\left[\frac{\gamma_*}{\eta_J} + ln(\Delta^2 - \epsilon_J)\right]^2\right) = \frac{\pi^{\frac{d}{2}}}{\Gamma(\frac{d}{2})}(\Delta^2)^{\frac{d}{2}-1}g(\Delta^2).$$

Setting $\epsilon_J = 0$, since it is a location parameter, yields

$$\frac{\Gamma(\frac{d}{2})}{\pi^{\frac{d}{2}}}\frac{\eta_J}{\sqrt{2\pi}}\exp\left(\frac{\gamma_*^2}{2}\right)(\Delta^2)^{-\frac{d}{2}}exp\left(-\eta_J\gamma_* ln(\Delta^2) - \frac{\eta_J^2}{2}ln(\Delta^2)^2\right) = cg(\Delta^2). \tag{3.23}$$

Using Equation (3.14) and only the terms that depend on Δ^2, and letting $z = \Delta^2$, $x = ln(z)$, and $dx = z^{-1}dz$ results in

$$C_d = \frac{\Gamma(\frac{d}{2})}{\pi^{\frac{d}{2}}} \left[\int_{-\infty}^{+\infty} e^{-(2+\gamma_* \eta_J)x - \frac{\eta_J^2}{2}x^2} dx \right]^{-1}, \tag{3.24}$$

which yields [23]

$$C_d = \frac{\Gamma(\frac{d}{2})}{\pi^{\frac{d+1}{2}}} \alpha_{SL}^{-1}, \tag{3.25}$$

where $\alpha_{SL} = \sqrt{p}\ exp\left(-\frac{q^2}{4p}\right)$, $q = (2 + \gamma_* \eta_J)$, and $p = \frac{\eta_J^2}{2}$.

Thus the final form of a multivariate EC density generated by the Johnson S_L-distributed MDs is

$$\mathbf{x}_{S_L} \sim \frac{\Gamma(\frac{d}{2})}{\pi^{\frac{d+1}{2}}} \alpha_{SL}^{-1} |\Sigma|^{-1/2} (\Delta^2)^{-d/2} \exp\left(-\eta_J \gamma_* ln(\Delta^2) - \frac{\eta_J^2}{2} ln(\Delta^2)^2 \right). \tag{3.26}$$

To verify the validity of this new multivariate form, synthetic two-dimensional multivariate data is created with the MDs from this data fit using a Johnson S_L model. Using the method described in [54] 10,000 two-dimensional data points are created that are governed by a mixture of multivariate t-distributions with degree of freedom parameters $\nu_{1,1} = 2$, $\nu_{1,2} = 2$, $\nu_{1,2} = 40$, and $\nu_{2,2} = 5$ and weighting coefficient $w = 0.8$. A mean data point is selected so that the joint density is centered at $x = 20, y = 20$. The MDs are calculated, and a Johnson S_L distribution is fit to the data. The resulting fit is shown in Figure 3.23.

Using the EC function for the multivariate t-density in Equation (3.21) with the parameters from the F-mixture fit to the MDs in Figure 3.23 and using the EC function derived here for a multivariate density resulting from the univariate Johnson S_L fit to the MDs, two-dimensional empirical densities models are created as shown in Figure 3.24. Notice the similarity between the derived EC model and the multivariate

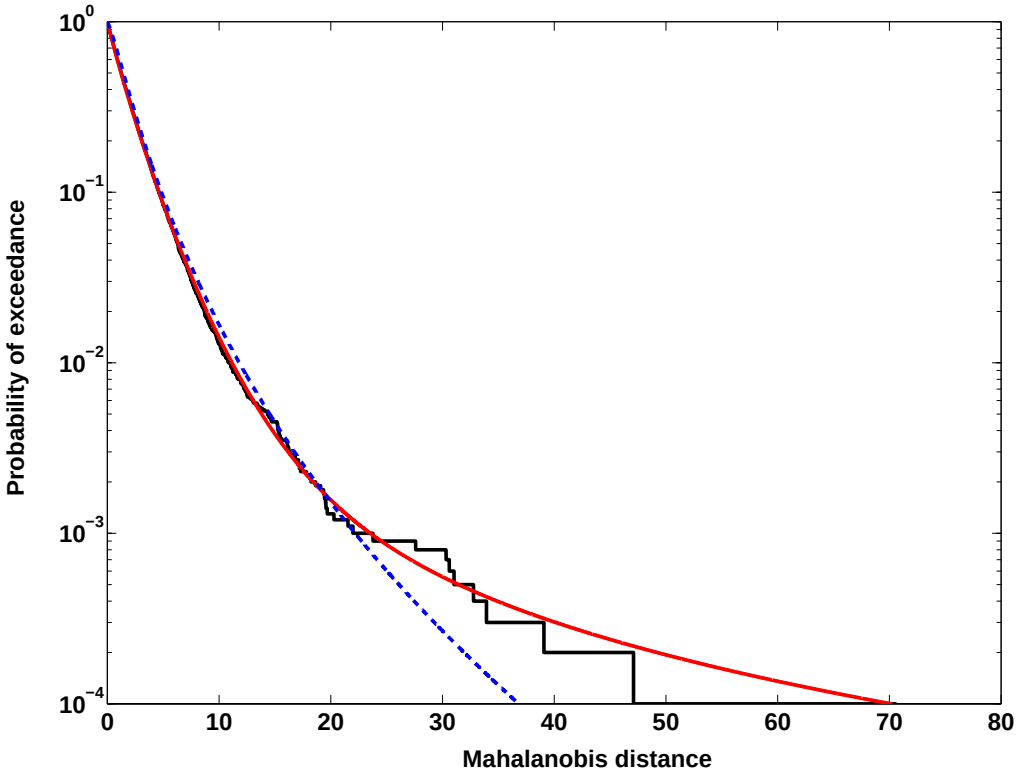
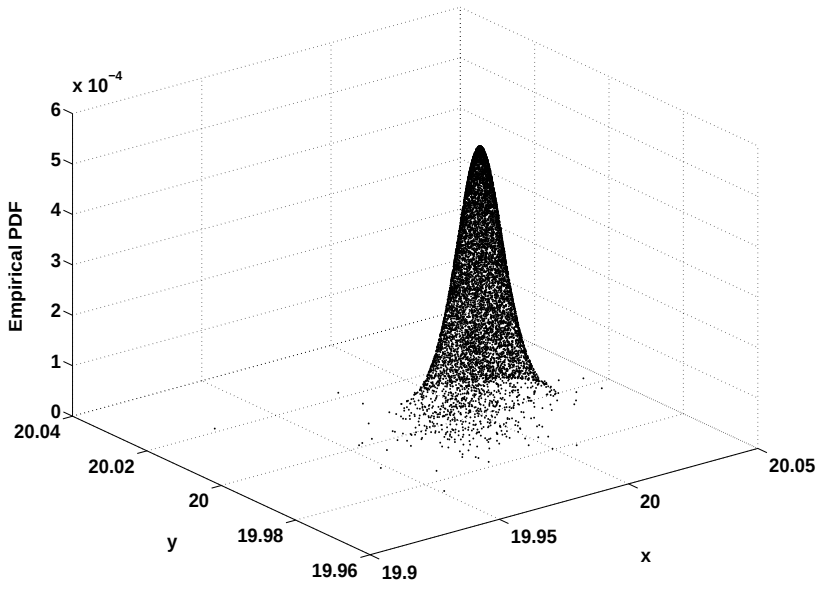


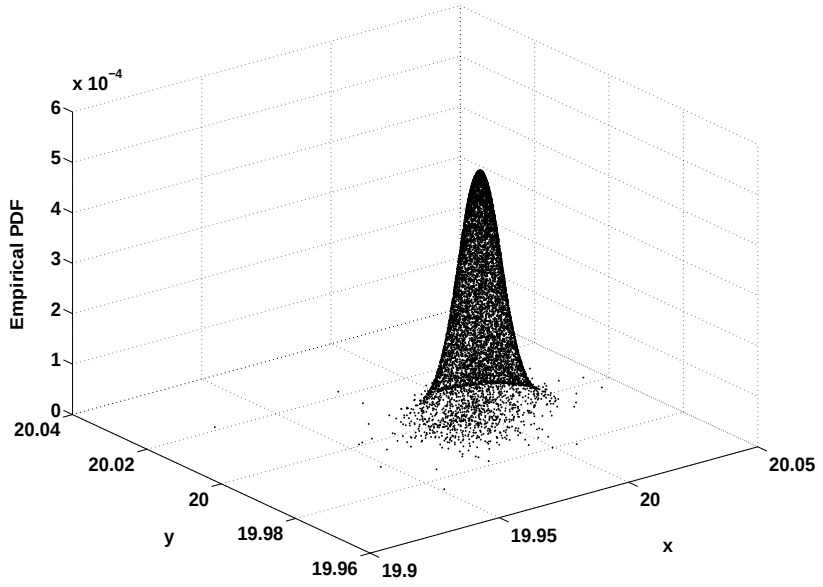
Figure 3.23: Here 10,000 two-dimensional data points are generated using a mixture of multivariate t-distributions with degree of freedom parameters $\nu_{1,1} = 2$, $\nu_{1,2} = 2$, $\nu_{1,2} = 40$, and $\nu_{2,2} = 5$ and weighting coefficient $w = 0.8$, where the joint density is centered at $x = 20, y = 20$. The MDs (dark line) are fit by a Johnson S_L distribution (dashed line) and a mixture of two F-distributions (light line)).

t-density. Based on the parameters and data in this example, a surface plot of the PDF for both probability density models is shown in Figure 3.25. Finally, the CDF of the derived EC density is shown in Figure 3.26.

3.5.3 Multivariate Synthetic HSI Data Generation. The form of Equation (3.26) is similar to many of the multivariate EC density functions given in the literature [8,21,54,64]. However, the research here does not provide a method for generating d-dimensional random variates characterized by the function in Equation (3.26). An expression for a multivariate EC density function is derived from the univariate Johnson S_L density function following the rigor outlined by EC theory. The result is a

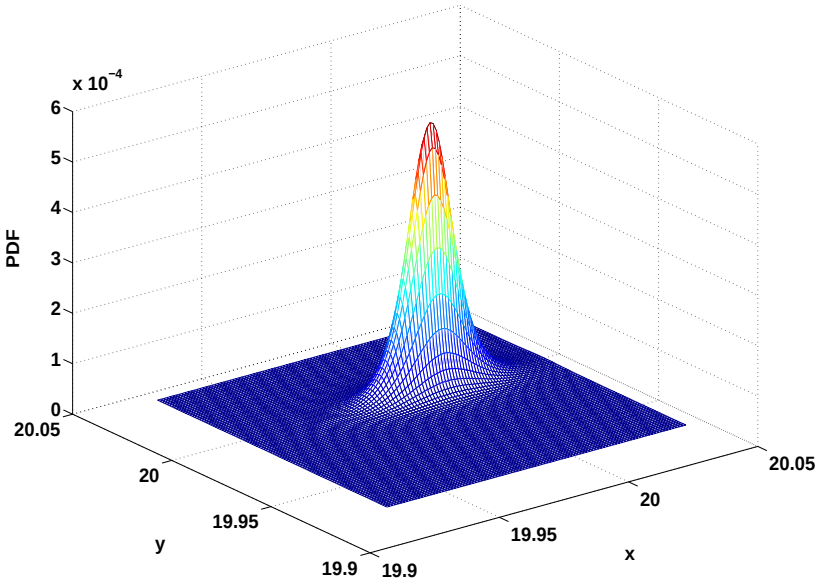


(a)

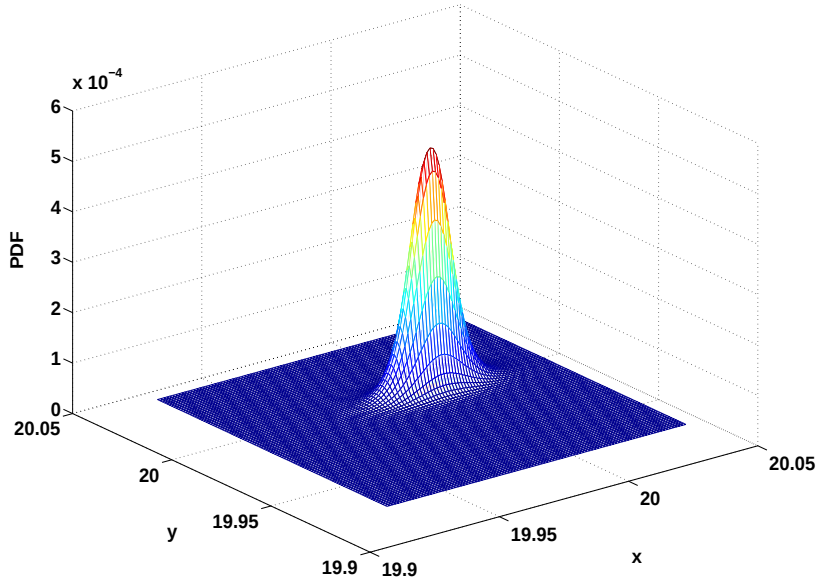


(b)

Figure 3.24: (a) A two-dimensional empirical density created using the EC multivariate t-density mixture with parameters from the univariate F-mixture that fit the MDs in Figure 3.23. (b) A two-dimensional empirical density created using the multivariate EC model derived here from the univariate Johnson S_L distribution. The Johnson parameters are from the Johnson S_L fit to the MDs in Figure 3.23. Notice the similarity between the two models.



(a)



(b)

Figure 3.25: (a) A two-dimensional probability density surface created using the EC multivariate t-density mixture with parameters from the univariate F-mixture that fits the MDs in Figure 3.23. (b) A two-dimensional probability density surface created using the multivariate EC model derived here from the univariate Johnson S_L distribution. The Johnson parameters are from the Johnson S_L fit to the MDs in Figure 3.23. Notice the similarity between the two models.

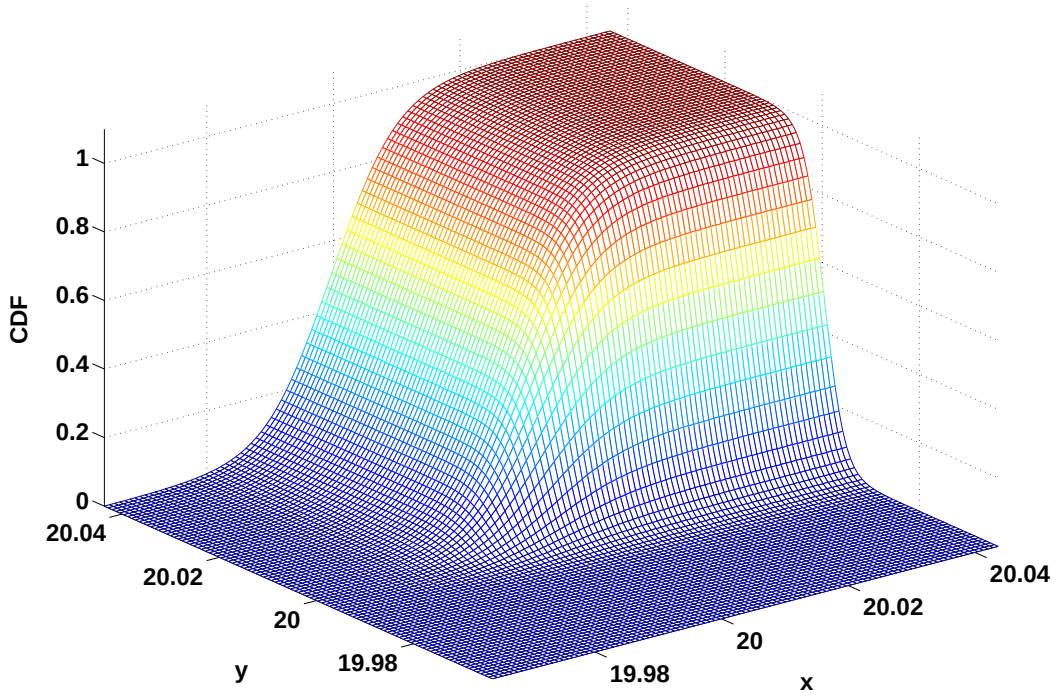


Figure 3.26: The CDF surface for the multivariate EC model derived here from the univariate Johnson S_L distribution that fits the MDs.

function similar to the expressions for multivariate Fisher-Bingham density functions, Kent density functions, asymmetric Laplace density functions, and density functions describing directional statistics [42, 57, 80, 81].

3.6 Summary

A robust Johnson S_L distribution model is identified and optimized for use with HSI data. Unlike the current technique for fitting MD distributions with mixtures of F-distributions, this method incorporates a correcting mechanism in the event of possible secondary processes. The method performs comparably to current techniques in modeling HSI MD distributions. However, the method is computationally more efficient than current methods, providing an improvement to current stochastic HSI data processing methods.

However, unlike the current technique, the multivariate EC density function derived from the univariate Johnson S_L density function does not recognize a known multivariate random variable generation algorithm to generate data that follow the EC model. Therefore, the next chapter investigates another method for modeling univariate MD data behavior, which is tractable to multivariate architectures that generate the observed MD distribution. Also, the new method uses only data from the tail of the MD distribution, as opposed to the entire data set.

IV. Tail-index Parameter Estimation Methods for GPD Models of MD Distributions

The Johnson S_L distribution has been identified and optimized for fitting MD distributions, thus addressing the goals of increasing computational efficiency and improving the robustness of fit. This chapter examines another method for modeling the heavy-tailed behavior of univariate distributions (with an emphasis on identifying the best model for HSI MD behavior), which is computationally efficient, amenable to optimization methods, and usable for synthetic data generation. A brief introduction to Extreme Value Theory (EVT) is given, then a standard method and a new method for identifying heavy-tailed behavior are presented. Finally, extreme value model parameter estimation techniques are evaluated, and the best method is identified for use with HSI MD data.

4.1 Extreme Value Statistics

Recent research in characterizing variability in HSI data shows that univariate HSI MD distributions (and therefore the distributions of multivariate pixel data) are described by heavy tail behavior [52, 55, 56]. Hence, the variability in HSI data is determined by the structure of the heavy tail of the MD distribution. More recently, it has been shown that the "heaviness" of tail behavior may be modeled by extreme value threshold methods [50, 52]. Specifically, an extreme value distribution model can be fit to the distribution of HSI MD data by analyzing only a small sample of the most extreme points in the tail of the distribution.

Of importance in HSI data exploitation is the capability to properly model the statistics of the data in order to develop robust post-processing techniques. Given the large size of HSI data, statistical models based on the entire data set are not computationally efficient. Therefore, the recent research in EVT models emphasizes determining robust statistical models given only a fraction of the data. EVT is a tool which draws information about the statistical nature of a process by analyzing only the extreme values of a data set. The basic probabilistic theory of extreme values has

been developed for many years. However, the statistical modeling of extreme data and relevant application in the field of HSI is comparably new.

4.2 Relevant Extreme Value Theory

The theory of extreme values provides a methodology whereby a fraction of samples in the tails of a data density may be used to obtain an accurate statistical model for the rest of the data. The ordered statistics (data points ordered from least to greatest in sequence, or vice-versa) sampled at the tails of the sequence have, asymptotically, the Generalized Extreme Value distribution (GEV) form [3, 9, 11]

$$G(z) = \exp\left\{-\left[1 + k\left(\frac{z-\mu}{\sigma}\right)\right]^{-1/k}\right\}, \tag{4.1}$$

where $-\infty < \mu < \infty$ is the location parameter, $\sigma > 0$ is the scale parameter, and $-\infty < k < \infty$ is the shape parameter, also known as the tail-index. For $k < 0$ the data distribution is governed by a short tail (bounded upper tail) density and has the form of the Weibull distribution. For $k = 0$ the limit behavior must be taken and the resulting expression is a Gumbel distribution, characterized by a thin tail density. For $k > 0$ the data distribution is governed by a heavy tail density and is classified as a Fréchet-Pareto distribution, in which case the mean is infinite for $k > 1$ and the variance is infinite for $k > 1/2$.

Also, of interest to this research is that heavy-tailed distributions belong to the domain of attraction of the Fréchet-Pareto distribution [3, 20]. For example, the F-distribution model of a data set is readily determined by the extreme value index k (tail-index). Therefore, modelling a data set with a GEV leads to known families of analytic forms of distributions in the domain of attraction of a GEV type. With the F-distribution form, it is simple, then, to derive the relationship between the distribution of MDs and the mutlivariate form of the distribution of the data set resulting in the MDs (see previous Chapter).

4.3 Discerning Heavy Tail Behavior

A well known method for describing tail behavior and a new method for quantifying the region of tail weight described by data quantiles are presented.

4.3.1 Mean Excess Function. A majority of extreme value statistical research is involved with heavy-tailed behavior of data sets. The heavy tail behavior of HSI data may be observed graphically through the use of exceedance plots (as in Chapters II and III). Traditionally, however, quantile plots have been employed in visual and quantitative methods for classifying distribution behavior.

For extreme value statistics, a more recent method for identifying heavy tail behavior is the sample mean excess function [3]. The mean excess function calculates the expected excesses over a specified threshold. Analytic mean excess functions for different distributions are plotted in Figure 4.1 using the parameters from Table 4.1. This information is adopted from [72] where $o(1)$ is a function of u whose limit is zero as $u \longrightarrow \infty$ and u is a threshold value. The exact values of the parameters are given in the Figure caption.

Table 4.1: Analytic mean excess functions for standard distributions.

Distribution	Mean excess function
Pareto	$\frac{\kappa+u}{\alpha-1},\ \alpha > 1$
Log-normal	$\frac{\sigma^2 u}{\ln u-\mu}\,(1+o(1))$
Weibull	$\frac{u^{1-\tau}}{c\tau}\,(1+o(1))$
Truncated Normal	$u^{-1}\,(1+o(1))$
Exponential	λ^{-1}

The mean excess over threshold value is an empirical estimate of the theoretical average above a given threshold u. It measures the expected value of the excess data given that exceedance has occurred and is

$$e_n(u) = \frac{\sum_{i=1}^{n}(X_{q(X_i>u)} - u)}{\sum_{i=1}^{n} q_{(X_i>u)}}, \tag{4.2}$$

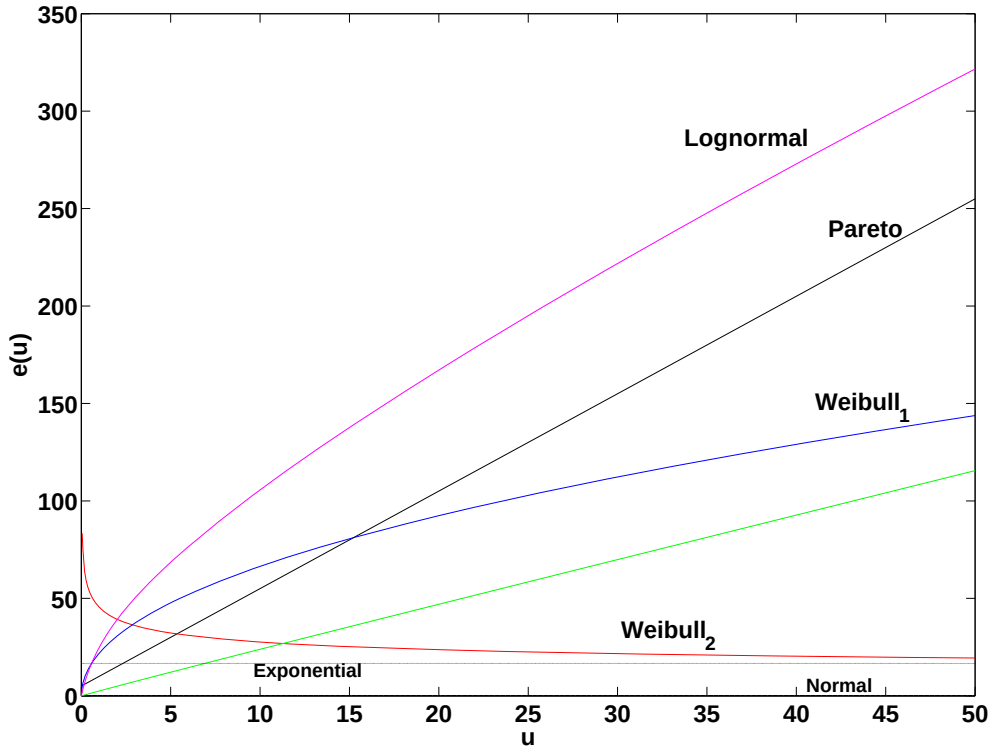


Figure 4.1: Mean excess functions for different distributions. The parameters, with respect to Figure **??** are: $\alpha = 1.2; k = 1; \tau_1 = 1.2; \tau_2 = 0.5; c = 0.1; \sigma = 5; mu = -2$. Notice the constant positive slope for the Pareto distribution (for all positive values of k, given $u\sigma > 0$). Notice the Lognormal and Weibull do not have a constant positive slope over the entire range of u.

where $q_{(X_i > u)}$ equals 1 for each sample value above the threshold and is 0 otherwise. Figure 4.2 shows an example of a sample mean excess function plot. If the plot shows an upward trend, then heavy tail behavior is indicated. Specifically, if the plot follows a linear positive slope for a region of u, then the data are distributed as a generalized Pareto distribution (GPD) with positive shape parameter k, which is the same k used in the GEV with the same behavior for positive, zero, and negative values. The other models showing positive slope can be represented as special cases of GPDs.

4.3.2 Quantile Ratio Function. Another technique for qualifying tail behavior and degree of "heaviness" uses distances between quantile values from a dis-

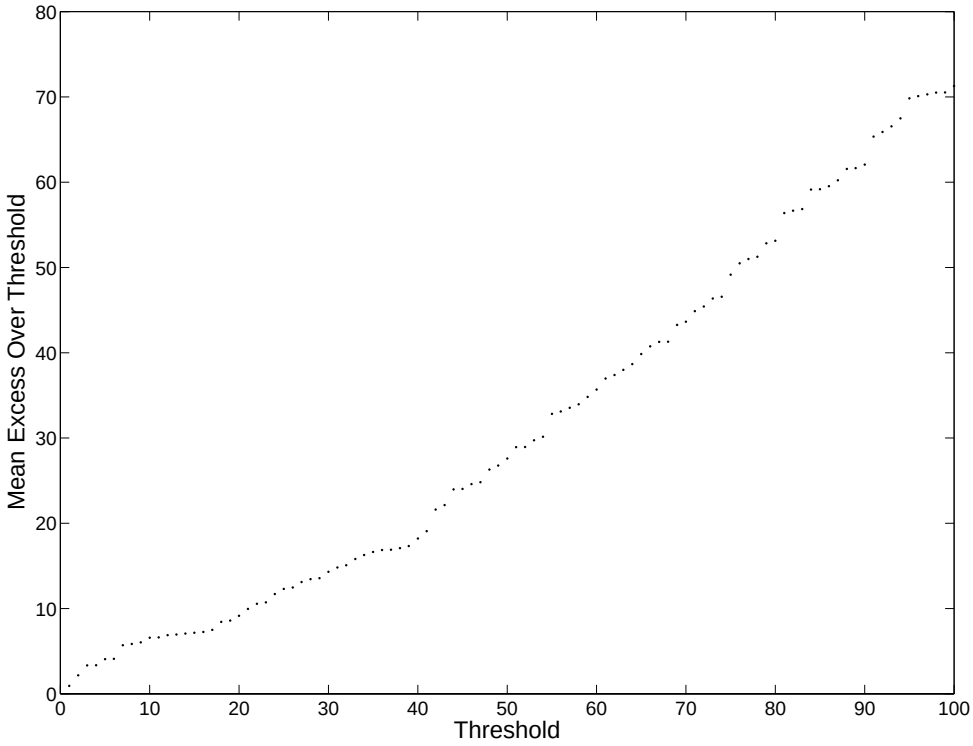


Figure 4.2: A plot of 100 sample mean excess points from 1000 random variables (RVs) generated from a GPD with a positive shape parameter $k = 1$, $\mu = 0$, and $\sigma = 1$. Notice the positive linear slope in the region between $u = 40$ and $u = 90$. The overall positive slope of this plot above the threshold values indicates heavy tail behavior.

tribution of data to determine the weight (heavy, light, or bounded) of the tail of a distribution. This method essentially compares the distances between two regions in the tail of a distribution to the distances of quantiles in the body of the data. The Empirical Quantile Ratio Function (EQRF) value used in discrimination is

$$\tilde{d} = \ln\left(\frac{t_1 t_2}{b^2}\right), \tag{4.3}$$

where t_1 is the distance between two empirical quantiles found in the "body" of the tail region, t_2 is the distance between two empirical quantiles in the extreme portion of the tail, and b is the distance between two empirical quantiles in the body of the

distribution. For example, in this work quantiles q_1 for $p = 0.002$, q_2 for $p = 0.85$, q_3 for $p = 0.97$, and q_4 for $p = 0.998$ are chosen such that $b = q_2 - q_1, t_1 = q_3 - q_2$, and $t_2 = q_4 - q_3$. Figure 4.3 shows the lengths of b, t_1, and t_2 for GP densities with different k values, and Figure 4.4 illustrates the behavior of $\tilde{d}$ given these quantiles for $-1 < k < 1$.

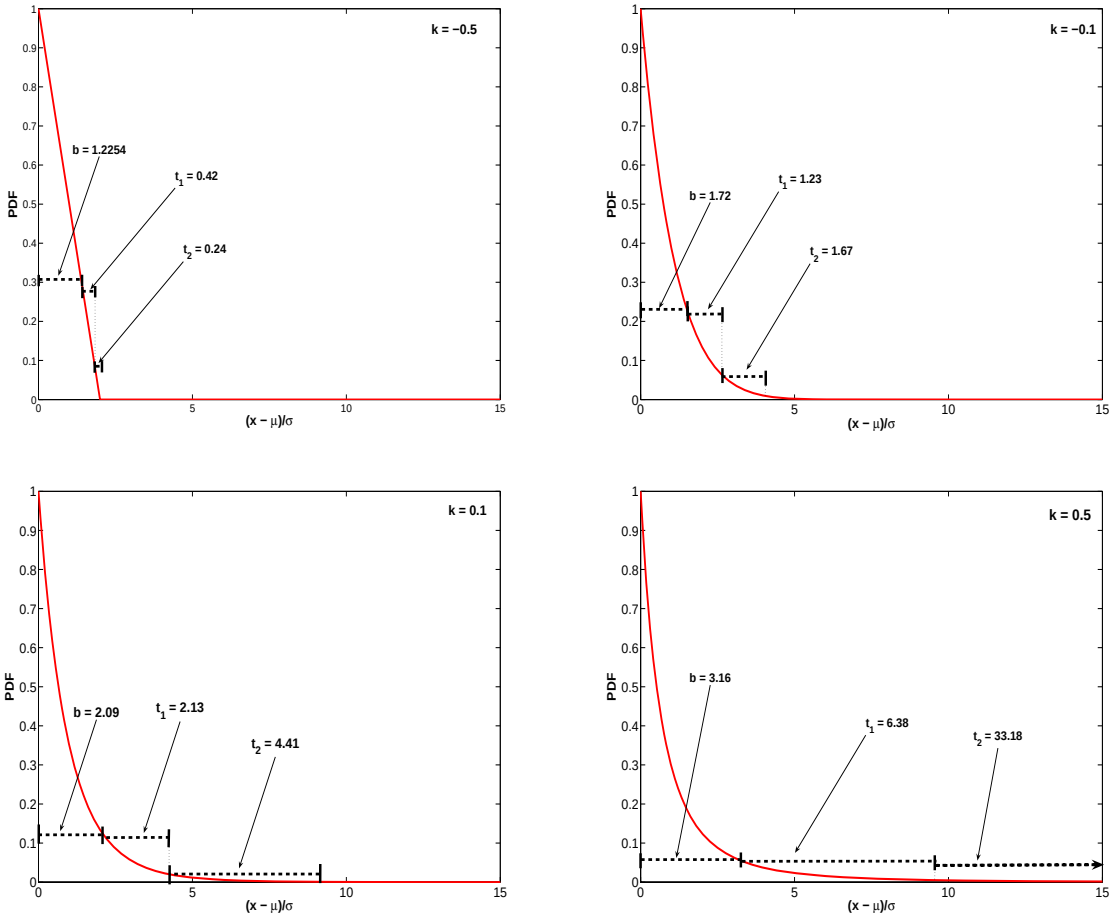


Figure 4.3: The lengths of b, t_1, and t_2 for GP densities with different k values. Notice the length of b does not change as dramatically as t_1 and t_2 when k increases. EQRF relies on this behavior for an initial guess at the value of k regardless of the size of the data sample.

Selecting different sets of quantiles representative of the tail regions and body of a distribution results in similar plots with an asymptotic empirical quantile ratio line increasing in a direction that is invariant for smaller data sets, because $\ln(t_1)$ and

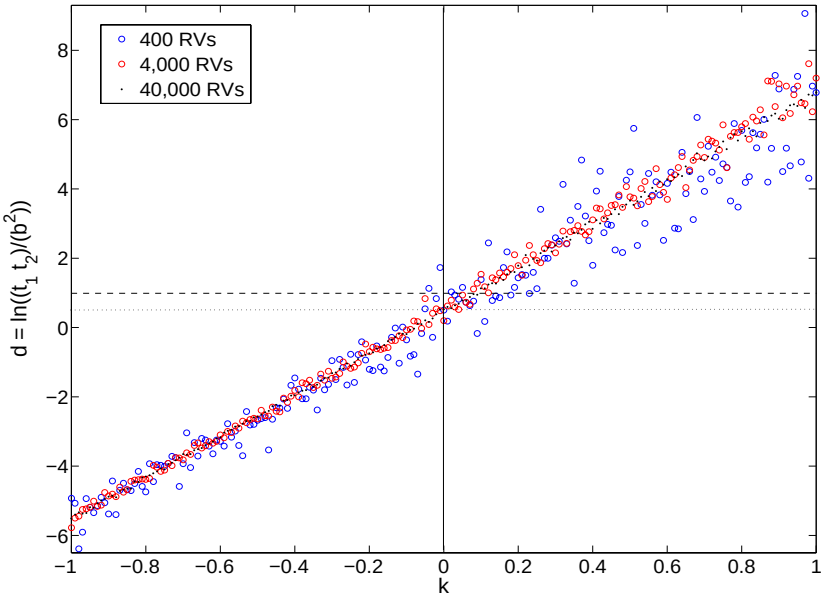


Figure 4.4: The quantile ratio function for $-1 < k < 1$ using 400, 4,000 and 40,000 RVs from a GPD with shape parameter k, $\mu = 0$, and $\sigma = 1$. Notice the increasing value of the plot as k increases. In particular, GPDs with positive k are recognized as those values of d above 0.5 (dotted horizontal line) for data sets of 40,000 points, above 1.0 (dashed horizontal line) for 4,000 points, and greater than 2.0 for smaller data sets, using the particular quantiles selected here. The slope may be different for other sets of quantiles used. Also, note that the slope of the asymptotic line defines the direction of the slope for smaller data sizes.

$\ln(t_2)$ increase much faster than $\ln(b^2)$ as k increases in Equation (4.3). Therefore, the change in d with respect to k has a positive slope when proper quantiles are selected for t_1, t_2, and b for a GPD. However, smaller data sets show greater variability in $\tilde{d}$ about the asymptotic line direction.

4.3.3 Application of Heavy Tail Qualification Methods. In this section the methods described in the previous section are applied to simulated data and to HSI data. For the first part, a data set of 10,000 RVs from a GPD with $k = 1.3$, $\mu = 150$, and $\sigma = 25$ is generated. This set is then analyzed for signs of heavy tail behavior in the mean excess function plot and empirical quantile ratio value. Figure 4.5 shows the mean excess function plot for the simulated data set and justifies using a GPD

to fit the data in a certain region. Using the quantiles mentioned in Section 2.2, the empirical quantile ratio value is 9.746, which indicates a positive tail index value of 1.5.

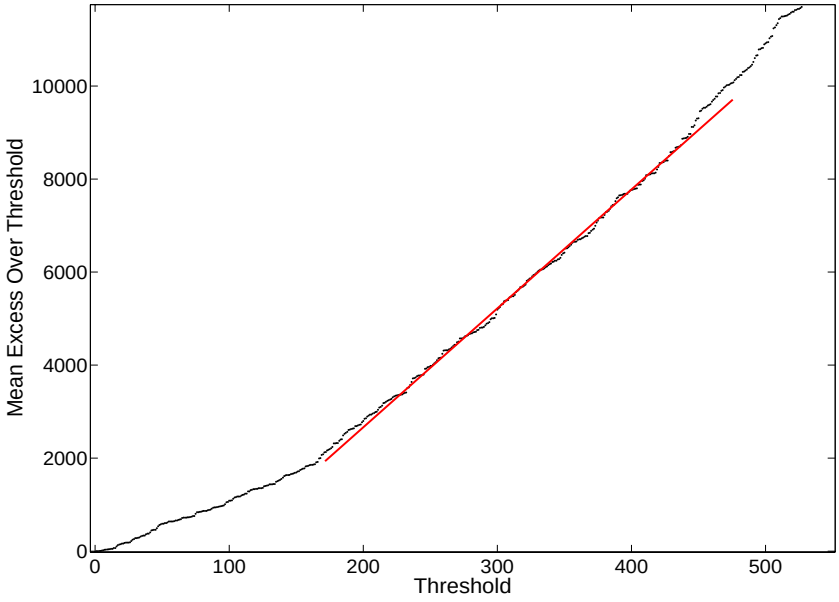


Figure 4.5: Mean excess function plot for 10,000 RVs from simulated GPD with $k = 1.3$, $\mu = 150$, and $\sigma = 25$. The plot shows a region where the mean excesses over a threshold follow an upward and linear trend, which is good visual evidence that the data are indeed heavy-tailed. The portion of the plot reasonably modeled by the straight line indicates that portion of the data above u follows a GPD with positive tail-index.

Next, a data set consisting of 17,000 MDs from a cluster of HSI data is also analyzed as above. The mean excess function plot is shown in Figure 4.6. The empirical quantile ratio value is 1.746, which suggests a positive k value of 0.27. Figure 4.7 shows an exceedance plot of the MDs and a GEV with $k = 0.27$, and indicates that the heavy tail parameter suggested by the empirical quantile ratio is reasonable.

Thus, two methods have been presented for qualifying the behavior of tail data in distributions. The mean excess function provides a visual method for identifying

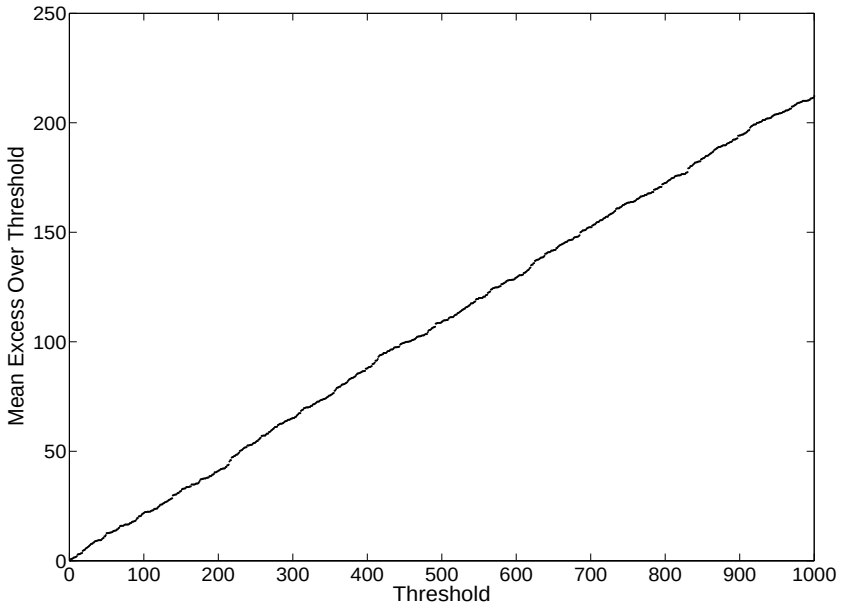


Figure 4.6: Mean excess function plot for 17,000 MDs from hyperspectral cluster data. The mean excesses over a threshold up to 1,000 follow an upward and linear trend. The visual evidence shows that the data are heavy-tailed out to this threshold.

heavy tail behavior. The empirical quantile ratio presents another method whereby a score is given, suggesting a degree of heavy tailed-ness.

However, neither of these methods is appropriate for estimating parameters for a GEV or GPD. They are somewhat contrived in that each method is constructed based on knowledge of tail behavior for GEVs and GPDs for a region of k values and a significant amount of data. However, they possess properties which develop a distinction between heavy-tailed data and the alternative (lighter or bounded tails). They may indicate how to apply other methods which exploit the statistical information in heavy tails for smaller samples of extreme values.

4.4 Exceedances Over High Thresholds

This section discusses methods for estimating the tail-index parameter k for these heavy-tailed processes. More precisely, based on a limiting property for extreme values, these methods are applied to samples above a threshold u. The sensi-

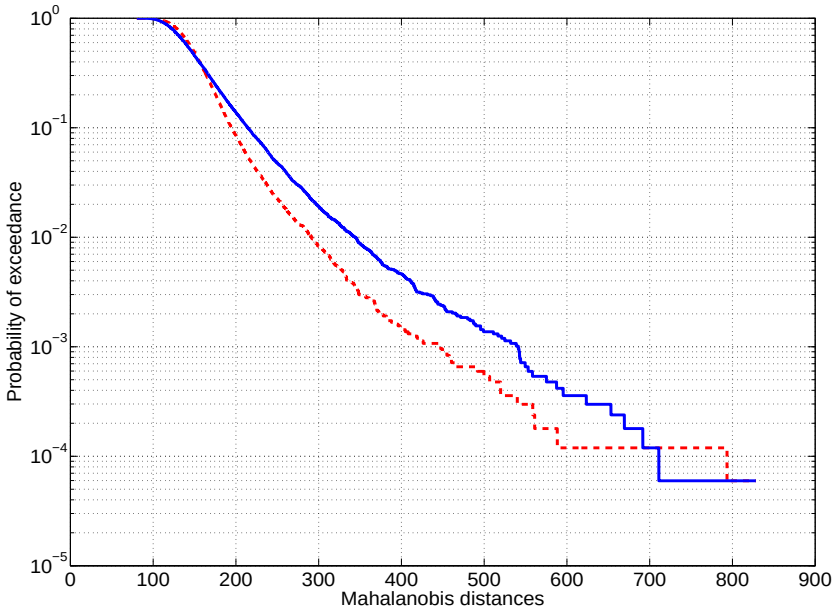


Figure 4.7: Exceedance plot of MDs from HSI data (dotted line) and GEV with $k = 0.27$, as approximated by the empirical quantile ratio of the data, $\mu = 150$, and $\sigma = 25$ (solid line).

tivity of each estimation method to threshold selection is examined for heavy-tailed distributions.

Let $X_1, X_2, \ldots X_n$ be a sequence of independent and identically distributed (i.i.d.) random variables (RVs), having a CDF F. Label the values of X_i that exceed some high threshold u as extreme events. Denote these values by $Y_1, \ldots, Y_m$, where $m \leq n$. Their CDF is [12]

$$F(y) = P\left(x > u + y | X > u\right) = \frac{1 - F(u+y)}{1 - F(u)}, \quad y > 0. \tag{4.4}$$

The CDF of the original values can be approximated using exceedances [11]. From

$$F_{x-u|X>u}(x) = \frac{F(x+u) - 1 + \theta}{\theta}, \tag{4.5}$$

where $\theta = 1 - F(u)$, the approximation for the original distribution is

$$F(x) = (1 - \theta) + \theta F_{x-u|X>u}(x - u), \quad x > u. \tag{4.6}$$

The GPD is used as a model for $F_{x-u|X>u}(x - u)$ that approximates the original random variable X [3].

If $X_{n,n} = \max\{X_1, \ldots, X_n\}$, then for large n [11, 12]

$$P(X_{n,n} \leq z) \approx G(z), \tag{4.7}$$

where $G(z)$ is the GEV distribution. For large values of u the CDF of $(X - u)$ given $X > u$ yields

$$H(y) = 1 - \left[1 - k\frac{(y - \mu)}{\sigma}\right]^{-\frac{1}{k}}, \quad y > 0, \tag{4.8}$$

which is the expression for the GP distribution function. The GPD arises as the limit distribution for the excesses over a threshold as the threshold increases toward the upper bound of the original distribution [62]. Also, if the threshold u is increased by an arbitrary amount, the GPD form remains unchanged, which defines the threshold stability property of the GPD. The probability density function (pdf) for the GPD is

$$f(y) = \frac{1}{\sigma}\left[1 + k\frac{(y - \mu)}{\sigma}\right]^{-\frac{1}{k-1}}, \quad k \neq 0 \tag{4.9}$$

$$= \frac{1}{\sigma}\exp[-(y - \mu)/\sigma], \qquad k = 0. \tag{4.10}$$

Figure 4.8 shows some examples of the GP pdf for different values of k.

As mentioned above, heavy tail data may be modeled by a GP density. Therefore using a Peaks Over Threshold (POT) method, one can estimate a shape parameter k that describes the the behavior of the entire distribution by simply examining a

smaller data set above a certain threshold. In the next sections, some appropriate estimation methods are analyzed for using POTs with GPDs to fit MD data.

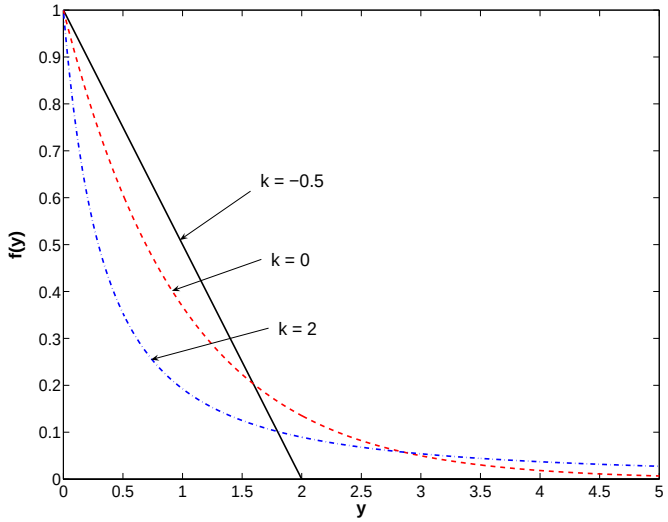


Figure 4.8: Examples of the GP density for different k values ($\mu = 0$ and $\sigma = 1$).

4.5 Estimation of Tail-index Parameter

The tail-index value k is the most important parameter for fitting a GPD to the distribution of HSI MDs. It defines the shape of the GPD and, hence, creates the greatest error in fit for small deviations in its value. The scale and location parameter are easily obtained once the shape of the function is determined. This section develops a general Bayesian method for estimating the k parameter. Methods derived from simplifications and approximations to the Bayesian estimator are then analyzed for use with HSI MD data sets.

4.5.1 Bayesian Estimation. Let data $D = \{Y_i, i = 1, 2, \ldots, n\}$ be i.i.d. samples from the density $p(y|\theta)$, where θ denotes the vector of parameters that define the form of the density. By Bayes' theorem

$$p(\theta|D) = \int_{\theta} p_y(y|\theta) p_\theta(\theta|D) d\theta, \tag{4.11}$$

where $p_\theta(\theta|D)$ is proportional to $p_D(D|\theta)P(\theta)$, and $p_D(D|\theta)$ is proportional to $\prod_{i=1}^{n} p_y(y_i|\theta)$. Therefore the posterior density is

$$p(\theta|D) = \int_\theta p_y(y|\theta) \prod_{i=1}^{n} p_y(y_i|\theta) P_\theta(\theta) d\theta, \tag{4.12}$$

which is a weighted average of possible densities.

In the Bayesian method, additional data tend to drive the solution from emphasis on the generalized prior (discussed in the next section) to emphasis on the data. That is, a small sample size results in a solution greatly influenced by the prior, while a larger data set overcomes this influence. Figure 4.9 depicts a schematic posterior pdf obtained from this method with maximum a posteriori (MAP) and minimum mean squared error (MMSE) estimators identified.

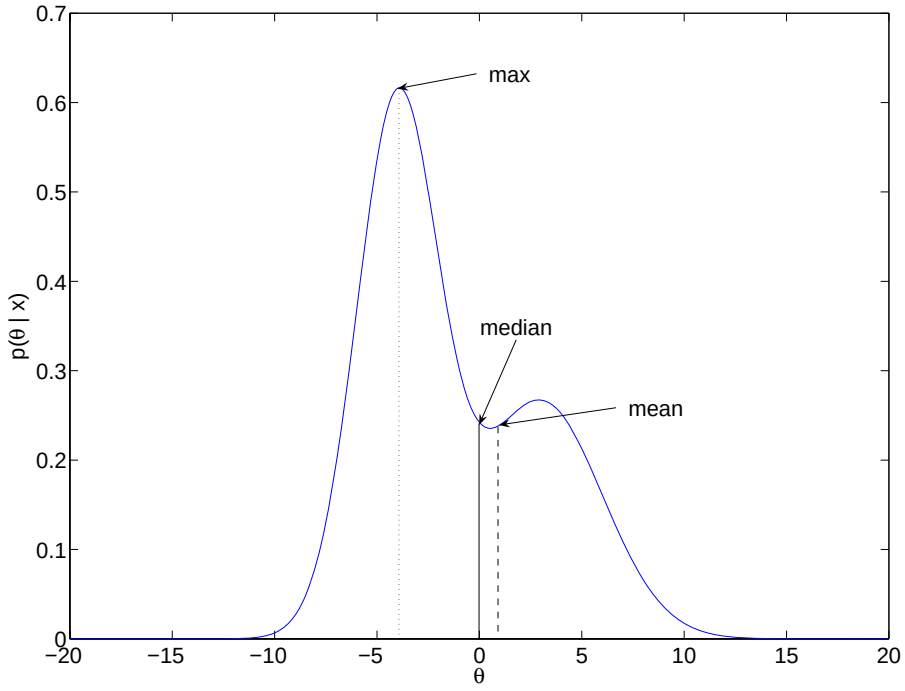


Figure 4.9: A schematic posterior pdf showing the locations of the maximum, mean, and median. The maximum is the MAP estimate, the mean is the MMSE estimate, and for a uniform prior, the maximum corresponds to the Maximum Likelihood (ML) estimate.

4.5.2 Selection of the Prior. A conjugate prior is a prior probability density which has the property that the posterior pdf has the same form. The GP density admits no conjugate prior density [3, 11], and Bayesian estimation typically selects a non-informative prior in this case. The objective is to select a prior for which information regarding parameter location is provided by the likelihood, while the parameter location envelope is provided by the prior [7].

Non-informative priors are typically from uniform, Jeffreys', or Gibbs sampling distributions, or from Markov Chain Monte Carlo (MCMC) techniques [12,13]. These priors are designed to form a minimal information framework for the prior, while keeping the prior within a realistic range specified by the data. Here a prior distribution form is selected based on knowledge of HSI data cluster MD distributions.

The tail-index k for a GP density fit to maxima exceeding a certain threshold tends to range from $-1 < k < M$, where M is some large positive value [52]. More realistically, the tail-index for GP models of MD distribution excesses from HSI data clusters range from $0 < k < \hat{M}$. In most applications involving GP models for a range of data sets $\hat{M}$ is rarely larger than 10 (actually, beyond $\hat{M} > 3$ the probability of encountering higher k values in real-world data decreases greatly). The range of k values for such data has a greater density in regions closer to zero and falls sharply beyond this region (i.e., the majority of MD distribution excesses rarely exhibit GP models with excessively large k values).

The gamma density is used here as the prior. It satisfies the above criteria and covers a wide range of shapes exhibited by the prior. Also, the maximum entropy criterion suggests the use of a gamma distributed prior: if the minimum of the prior is greater than zero and the p^{th} quantile equals some finite value, then the criterion specifically prescribes use of a gamma distributed prior [47]. The gamma pdf is

$$f(x) = \frac{\beta}{\Gamma(\alpha)}(\beta x)^{\alpha-1}e^{-\beta x} \qquad x \geq 0, \tag{4.13}$$

where β and α are positive scale and shape parameters, respectively. For $\alpha = 1$ the gamma density is equivalent to the exponential density, and for $\alpha = \frac{h}{2}$, where h is an integer and $\beta = 2$, it is a chi-square density with h degrees of freedom. Different choices for β and α result in uniform, beta, and Pearson Type VI densities, among others [36]. Figures 4.10 and 4.11 show gamma density functions for various α values with $\beta = 1$ and $\beta = 2$.

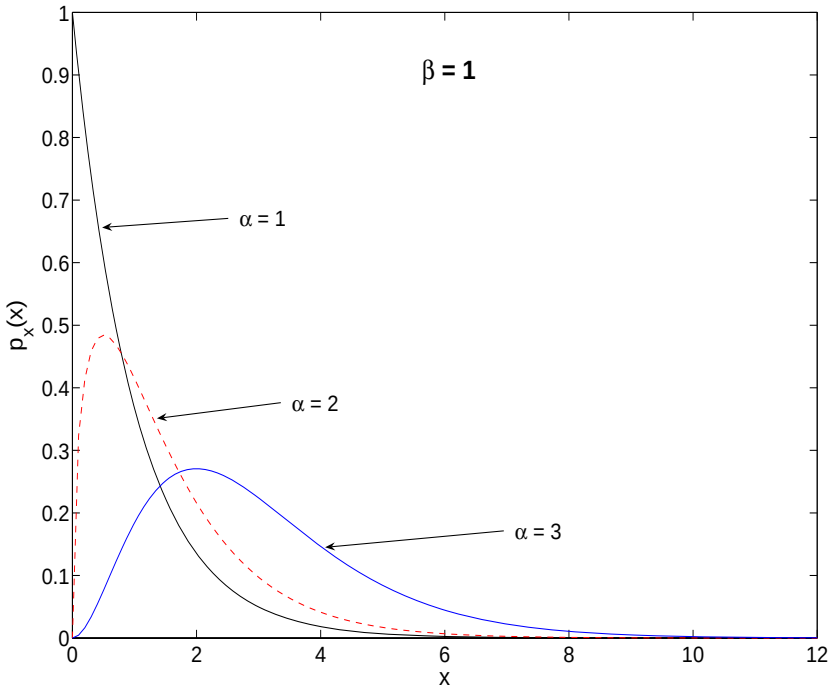


Figure 4.10: Gamma density functions for different α and with $\beta = 1$.

4.5.3 Application of Bayesian Estimation Method. In the following a set of 1,000 points is generated from the GP pdf with $\mu = 0$, $\sigma = 1$, $k = 1.0$, and, initially, $u = 900$ POT samples are used to develop a posterior density for the desired k parameter. A large number of POTs are used for this analysis to obtain a well-developed posterior distribution and to avoid having the prior overwhelm the influence of the data. In this case θ in Equations (4.11) and (4.12) is the k parameter. In some literature θ is a vector with parameters k and σ comprising the vector elements [13],

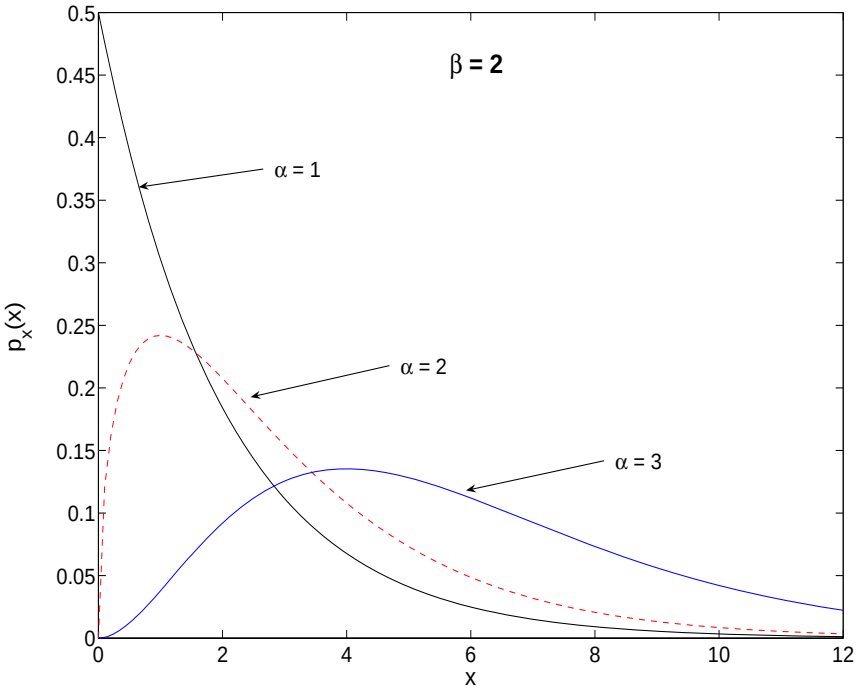


Figure 4.11: Gamma density functions for different α and with $\beta = 2$.

[66]. However, here the only concern is estimating k, because in general it determines the behavior of the extremes. Also, in many multi-parameter estimation techniques the remaining parameters are easily obtained once k is estimated.

A prior gamma pdf is specified with parameters $\alpha = 1.5$ and $\beta = 1$, and 2000 candidate values are drawn from this prior distribution in the range $0 < k < 10$ and then used in Equation (4.12) to generate Figure 4.12, which shows the shape of the prior density on k and the posterior density as a result of the Bayesian process. Notice the maximum is very close to the actual value of $k = 1.0$ and that the shape is highly peaked towards the correct value, falling off dramatically away from the true value. This effect is due to the influence of the prior and the size of the sample. Knowing that the majority of k values for heavy-tailed distributions are generally less than $k = 3$ and more likely for $0 < k < 1.0$, the shape of the prior is selected to take advantage of this knowledge. Also, having a large data sample insures that the data

likelihood will avoid being biased by the prior. For example, in Figure 4.13 only 100 POTs are used. Notice how the posterior becomes wider and peaks away from the actual value of $k = 1.0$ so that the prior biases the estimator even though the prior is selected such that it represents the behavior of k values for heavy-tailed distributions. However, in Figure 4.14 these same 100 data points are used with a prior that peaks away from the actual value of $k = 1.0$ and displays an incorrect pattern for heavy-tailed distributions. Obviously, the estimate becomes even more biased towards an incorrect value.

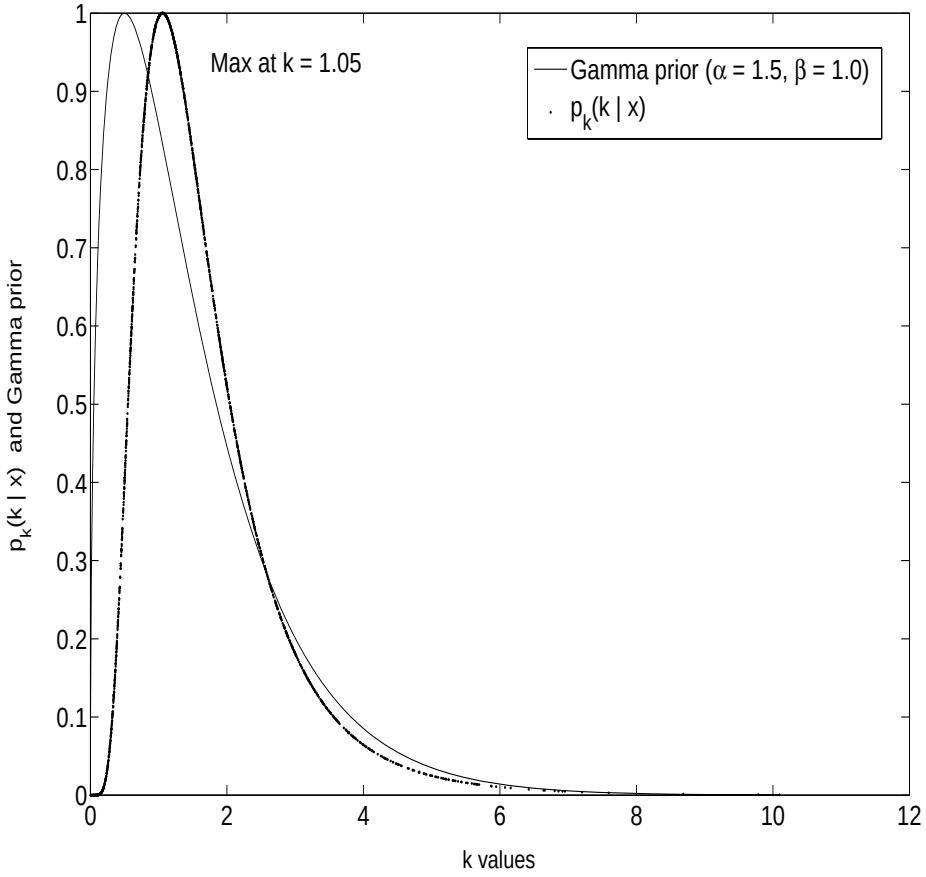


Figure 4.12: Prior and posterior density function shape for Bayesian estimation of k given 900 POTs selected from a GP distributed data set with $k = 1.0$, $\mu = 0$, and $\sigma = 1$. Notice the maximum of the posterior is close to the actual value of k.

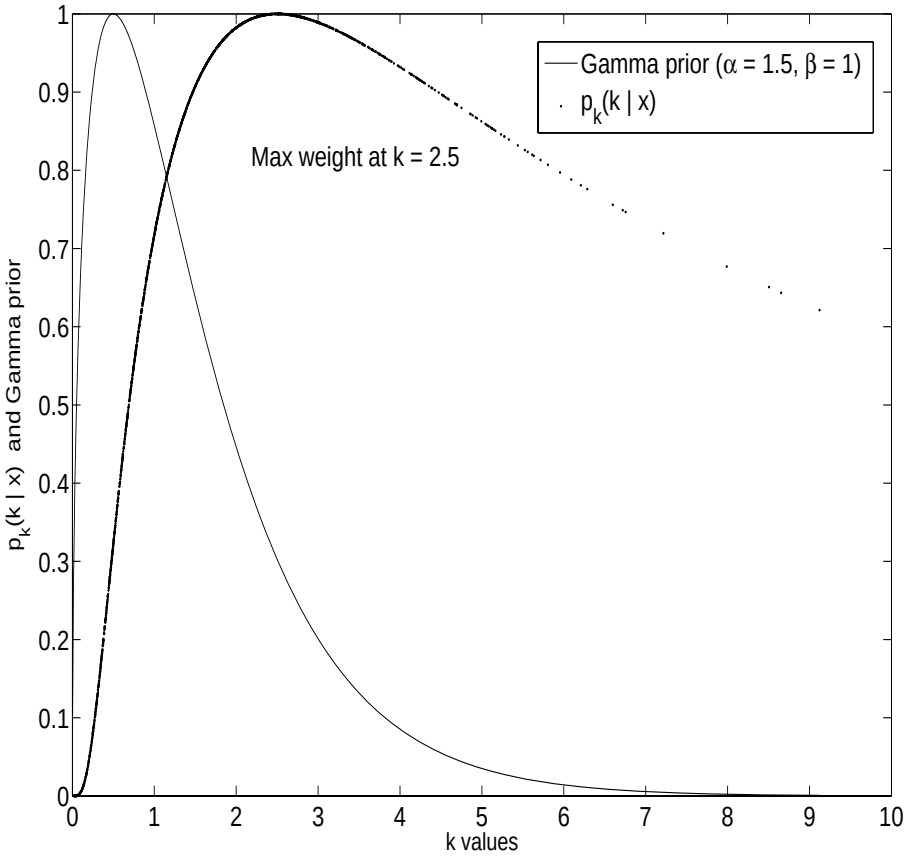


Figure 4.13: Prior and posterior density function shape for Bayesian estimation of k given 100 POTs selected from a GP distributed data set with $k = 1.0$, $\mu = 0$, and $\sigma = 1$. Notice the maximum of the posterior deviates from the actual value of k. In this case there are too few data samples to overcome the influence of the prior.

Aside from the detrimental effects of selecting an inappropriate shape for the prior and too few data samples, the advantage of this type of estimation is that metrics are inherent in the posterior probabilities. Using the values of the posterior as weights on possible GP densities to fit the data set under analysis, a surface of GP models appropriate to fitting the data can be created as in the lower left quadrant in Figure 4.15. Taking a vertical cut of this surface at any point on the x-axis generates a pdf of k given the data. The peak indicates the most probable value. Similar simulations show that with fewer data points the prior overwhelms the data and a broad peak

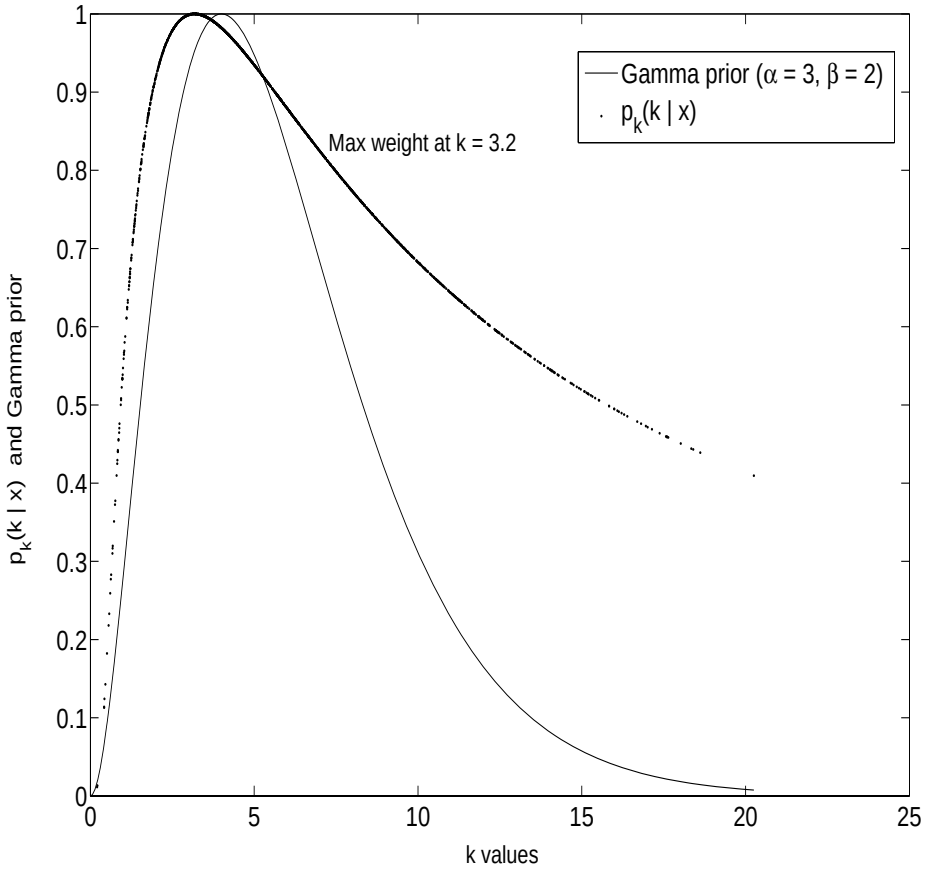


Figure 4.14: Prior and posterior density function shape for Bayesian estimation of k given 100 POTs selected from a GP distributed data set with $k = 1.0$, $\mu = 0$, and $\sigma = 1$. Here, the prior has change to a density not representative of the behavior of k for heavy-tailed distributions. Notice the performance of the estimator degrades even more so.

results. The lower right quadrant of Figure 4.15 is a two dimensional representation of this surface.

Also of interest is the feasibility of using the Bayesian estimation method for certain regions of k. As mentioned, it is particularly simple to select the Gamma prior for a high density in the region $k < 2.0$. However, selecting a prior for heavy-tailed distributions with larger values of k may be difficult because the data likelihood of the largest extreme values overwhelms the likelihood scores of the rest of the sample. Selecting a proper prior indicative of this type of behavior is difficult. Therefore, the

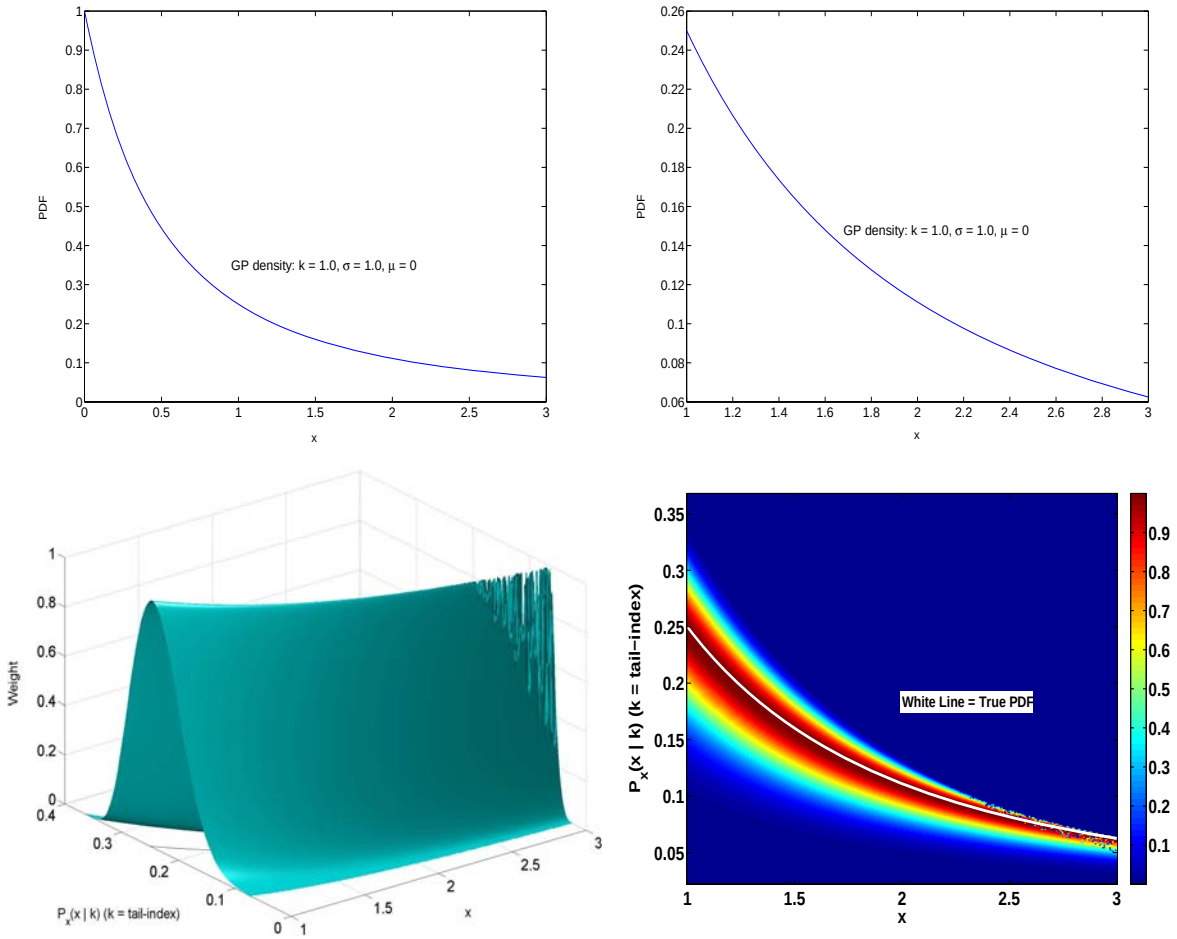


Figure 4.15: Upper left: GP density of RVs selected for this Bayesian estimation analysis. Upper right: Same GP density emphasized to show the region $1 < x < 3$. Lower Left: The surface resulting from GP densities at each tail-index value (k) weighted by the corresponding posterior probability value $P_k(k|x)$ in the region $1 < x < 3$. The peak of the surface occurs at $k = 1.05$, which corresponds to the MAP estimate, from 900 samples taken from a GPD with actual $k = 1.0$. Lower Right: A two dimensional representation of the surface.

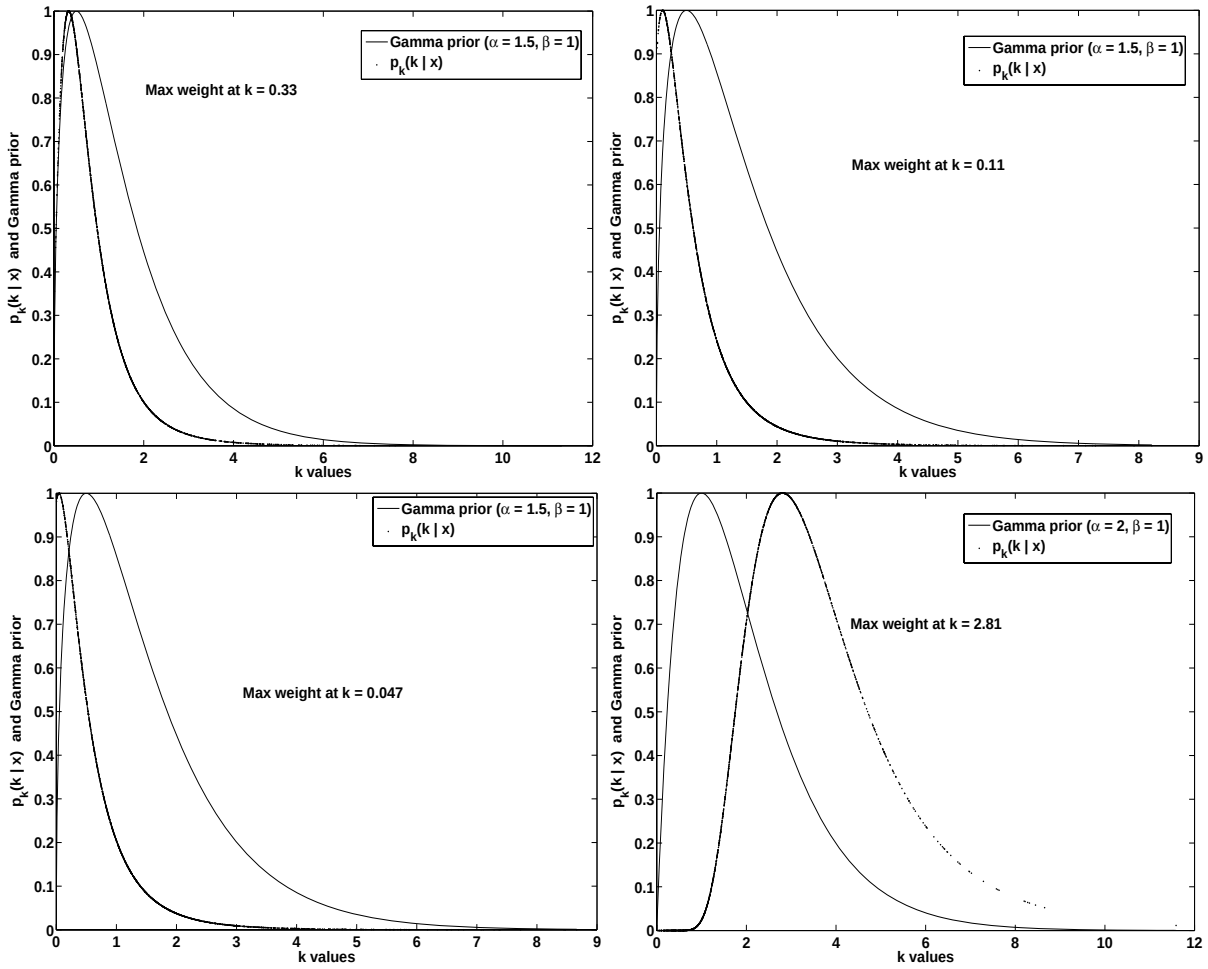


Figure 4.16: Upper left: Posterior k density for GP with actual $k = 0.3$. Upper right: Posterior k density for GP with actual $k = 0.1$. Lower Left: Posterior k density for GP with actual $k = 0.05$. Lower Right: Posterior k density for GP with actual $k = 2.3$. Notice how the posterior density is localized and highly peaked at the actual k value for the region $k < 2.0$. In the lower right plot, actual $k = 2.3$ and the posterior becomes less peaked and deviates from the true value. As k increases in the region $k > 2.0$ the posterior becomes flatter and deviates from the true value. In all cases $\mu = 0$ and $\sigma = 1$.

optimal region for using the Bayesian estimation process for heavy-tailed distributions is $k < 2.0$. This conclusion is illustrated in Figure 4.16.

4.6 Bayesian Estimation Method for GPD Parameter Density

Since application of the traditional Bayesian estimation approach is limited to a certain range of k, it is not identified as the proper estimation routine for use in esti-

mating k to fit GPD models to MD distributions. Point estimators (generalizations of the traditional Bayesian estimation method) are investigated as the alternate. However, the utility of traditional Bayesian estimation is the generation of a parameter density.

Using Bayesian parameter estimation with Parzen windows, a new technique for determining the density of the parameter k is developed as a result of this research. The Bayesian estimator outputs discrete values of posterior density information for k given a prior. Then, Parzen windows density estimation with a gamma density kernel is applied to generate a smooth posterior density profile.

A demonstration of this method, 10,000 random variables (RVs) are generated from a GPD with parameters selected to mimic HSI MD distribution behavior [59]. The parameter values are: tail index $k = 0.2$, scale parameter $\sigma = 26$, and position parameter $\mu = 125$. A plot of the empirical probability of exceedance is shown in Figure 4.17.

Estimation of the parameter μ given a uniform prior does not give robust results under the Bayesian framework, because as μ increases the probability density values of larger tail data points increase for a GP density. The likelihood is a product of probability density values for the data given a parameter set. Therefore, the Bayesian estimator consistently selects higher μ values, as these values generate larger likelihood values.

In much the same manner the Bayesian estimator selects the maximum σ value, because a larger σ value corresponds to larger pdf values and hence larger likelihood scores. Thus a non-uniform prior must be specified for the σ parameter to obtain robust Bayesian estimates. Since ML estimation for GPD models requires a uniform prior, only the k parameter is considered here, as it is the only parameter for which likelihood score is not adversely affected by increasing parameter magnitude. Also, note that once k is estimated the posterior density follows the shape of the posterior

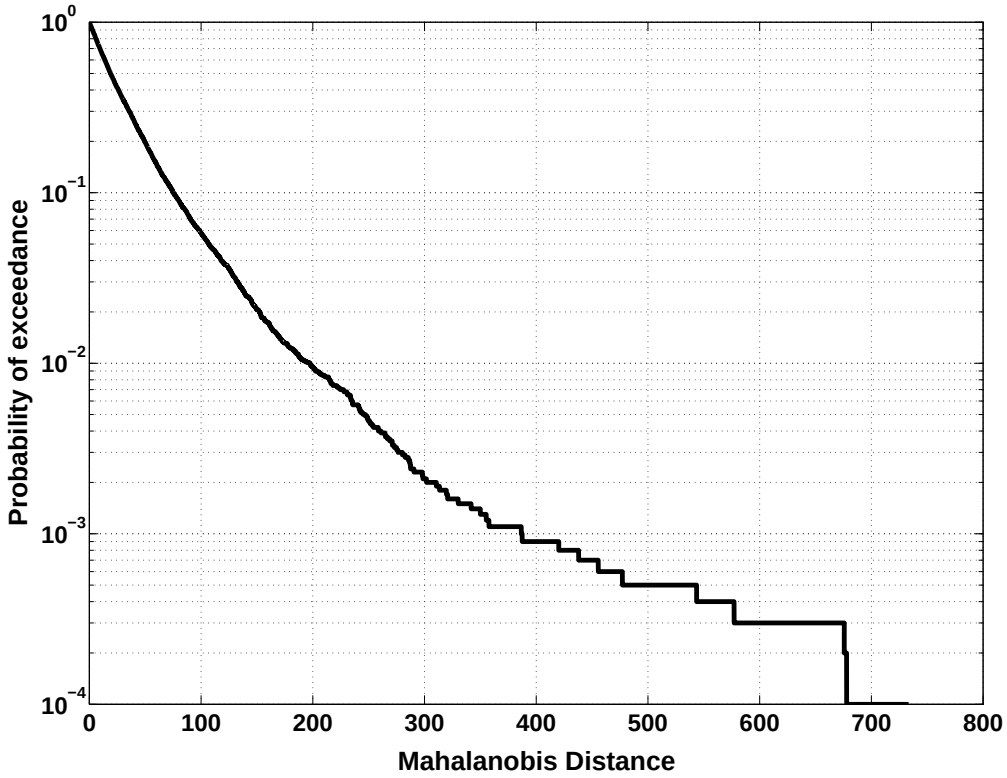


Figure 4.17: Empirical probability of exceedance plot for 10,000 values generated from a generalized Pareto distribution with parameters $k = 0.2$, $\sigma = 26$, and $\mu = 125$. These parameters are similar to those estimated from GPD models fit to vegetative HSI MD distributions [59].

density for k, since σ and μ only change the magnitude and location of the posterior density and do not affect shape.

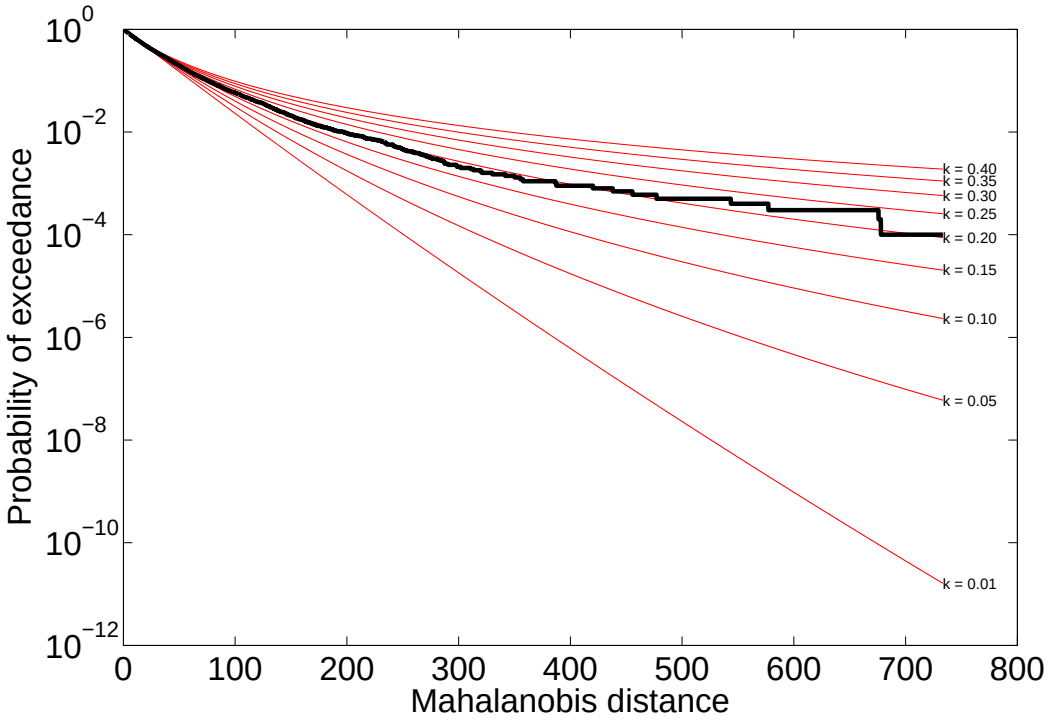


Figure 4.18: Probability of exceedance plot for GP random variables generated with parameters $k = 0.2$, $\sigma = 26$, and $\mu = 0$ (thick line), and nine models of GPDs spanning the range $0.01 < k < 0.4$ in nine increments are shown (thin lines).

Initially, nine increments of k in the interval $0.01 < k < 0.4$ are used as a uniform prior for estimating the tail-index of GP random variables generated with the actual values of $k = 0.2$, $\sigma = 26$, and $\mu = 0$. The probability of exceedance is shown in Figure 4.18 along with the candidate GPD models for the nine different values of k. Referring to the separate curves for k in the Figure, a σ value shifts the curve up or down (depending on magnitude) a small increment, and μ shifts the curve up or down in smaller increments.

The results of the Bayesian estimation are given in Figure 4.19, which shows a vertical slice through Figure 4.18 at the last value. Thus it is a profile of the posterior density for probability of exceedance at the largest MD. Notice the shape of the

posterior density for probability of exceedance of GPD models available for fitting the generated data, which is directly related to the logarithmic scale of the x-axis. Taking a vertical cut of the exceedances at smaller MD values, where a logarithmic scale is no longer necessary to represent the exceedance value range, results in a posterior shape of the probability of exceedance that follows the shape of the posterior density of k values.

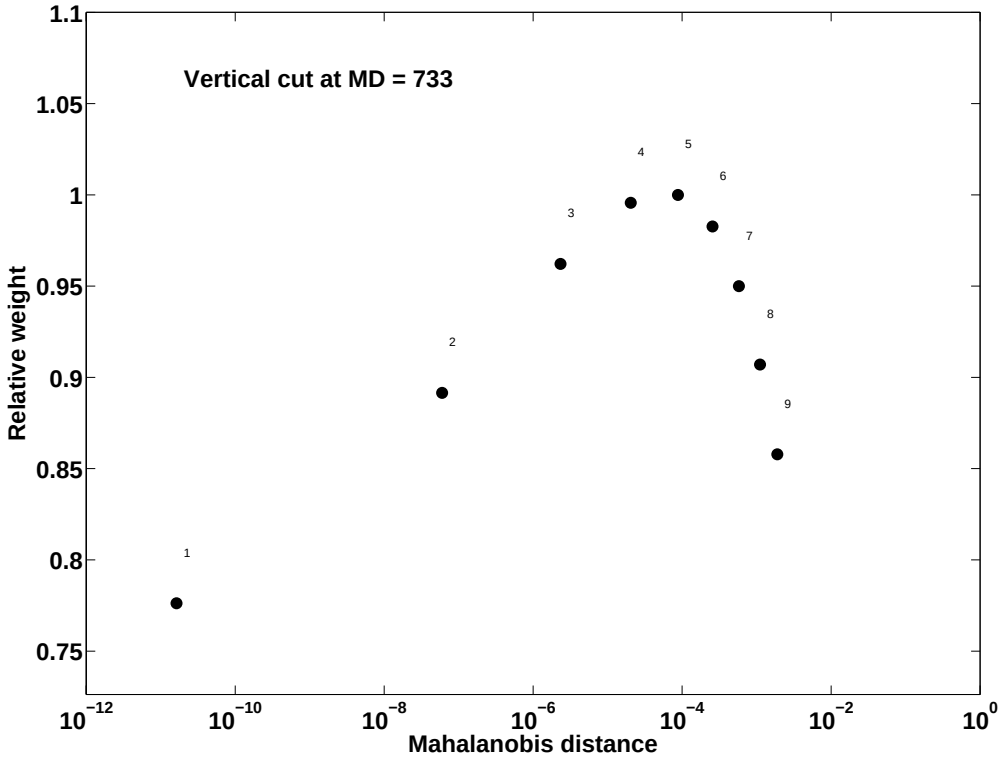


Figure 4.19: A vertical cut at MD = 733 shows a rudimentary profile of the posterior density for the nine models of GPDs that fit the data in Figure 4.18. The "Relative weight" on the y-axis is the value obtained from the likelihood calculation for each model given the data.

The posterior density for k may be obtained by Parzen window density estimation [19]. For this example, since for heavy-tailed behavior $k > 0$ is required with an unbounded upper limit, a gamma density is used for the Parzen window kernel. The gamma density satisfies the above criteria and covers a wide range of shapes. Figure 4.20 shows the Parzen window estimate of the posterior density for k with a

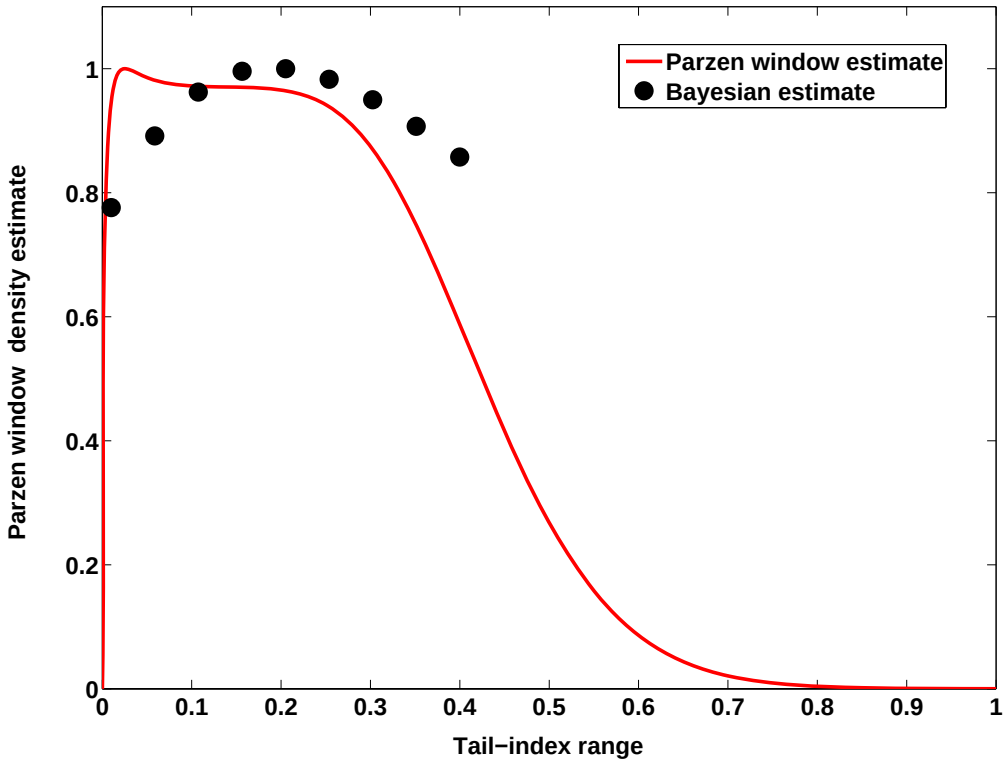


Figure 4.20: Nine posterior weights for the k values used in the Bayesian estimation. The shape of the density for k based on these weights is estimated using Parzen windows with a gamma pdf as the kernel. Here the marginally unimodal estimate is shown. Notice the higher peak at the left. This peak is indicative of the shape of the window. In this case the peak is created from the superposition of the gamma function window on the left-most data point and the window on the next point Adjusting the variance on the windows creates different shapes for the density.

marginally unimodal shape. The variance on the gamma density kernels is adjusted such that the shape of the density is bimodal, then the variance is increased until the density becomes marginally unimodal. In Fig. 4.21 variance is increased until the least-squares best fit estimate is obtained. Notice that this density has a maximum close to $k = 0.2$, the actual tail-index value.

The results demonstrated here are important because they provide a capability for extracting metrics on the estimated tail-index value. For ML estimation, confidence intervals and further metrics may be derived from asymptotic properties.

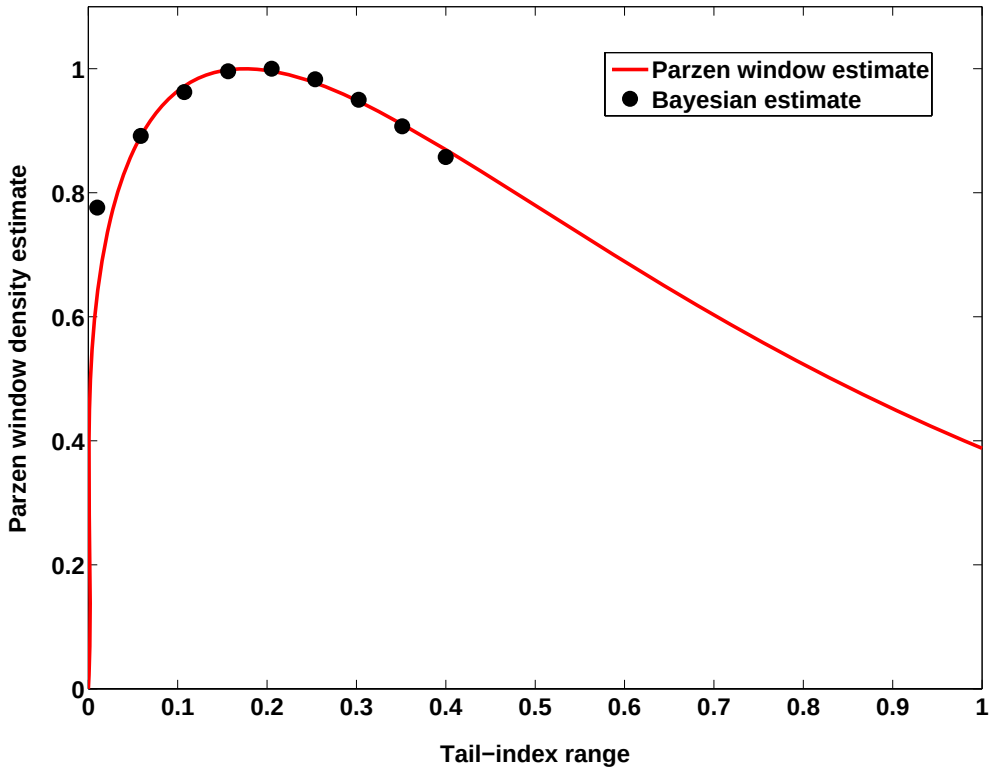


Figure 4.21: The best fit (in the least squared error sense) Parzen window estimate of the shape of the density of k given the nine points estimated using the Bayesian estimation method. Compare to the marginally unimodal case in Figure 4.20. Notice that the maximum of the estimated density is close to the true value of $k = 0.2$.

However, for real-world and real-time data, sample sizes rarely approach the asymptotic limit, and a method that is effective with the data available is necessary. The method described here demonstrates the development of a density that estimates the tail-index parameter of a data set (similar to HSI MD data) using only nine inputs.

Using this estimated posterior density obtained from an initial limited input set allows for fast and efficient processing of HSI data using the GPD model for MD distributions, for which metrics are obtained without asymptotic assumptions. Furthermore, as the method assumes nothing beyond the data, more realistic and meaningful conclusions (with respect to the information contained in the data) may

be formed. This capability is important for properly characterizing the statistics of a suite of images in which the data changes from one image to another.

4.7 Other Estimators (Special Cases of the Bayesian Estimator)

Having developed a method for determining the posterior pdf generation capability, the best method for estimating the k parameter value (based on computational efficiency and not limited to a specific range of tail-index values) is selected upon evaluating a number of point estimators. The following estimators are developed from approximations and/or simplifications of Bayesian estimation. For example, where Bayesian estimation produces a pdf for the estimate, the following methods are only point estimates. They often correspond to a single point on the posterior pdf, and further information is required to develop their variance and higher order moments.

4.7.1 Maximum Likelihood Estimation of GPD Parameters. The maximum likelihood estimator is a special case of the Bayesian estimator in which the prior is assumed to be constant (uniform or flat) as a function of parameter values, and the maximum of the posterior density is the single point estimate. Given a threshold $u = X_{n-u,n}$ at one of the sample points with exceedances above the threshold denoted $Y_{j,u} = X_{n-u+j,n} - X_{n-u,n}$ for $j = 1, 2, \ldots, u$ [3], the log-likelihood function for the N_U excess RVs are (based on the GPD model)

$$\ell(k, \sigma|Y) = -N_u \ln \sigma - \left(\frac{1}{k} + 1\right) \sum_{i=1}^{N_u} \ln\left(1 + \frac{kY_i}{\sigma}\right), \quad k > -\frac{\sigma}{Y_i}, \quad i = 1, \ldots, N_u. \tag{4.14}$$

For $k = 0$ the log-likelihood function is

$$\ell(0, \sigma|Y) = -N_u \ln \sigma - \frac{1}{\sigma} \sum_{i=1}^{N_u} Y_i. \tag{4.15}$$

With $\eta = k/\sigma$, Equation (4.14) becomes [25]

$$\ell(k, \eta|Y) = -N_u \ln k + N_u \ln \eta - \left(\frac{1}{k+1}\right) \sum_{i=1}^{N_u} \ln\left(1 + \tau Y_i\right). \tag{4.16}$$

The single parameter log-likelihood is then

$$\ell(\eta|Y) = -N_u - \sum_{i=1}^{N_u} \ln\left(1 - \eta Y_i\right) - N_u \ln\left[-\frac{1}{N_u} \sum_{i=1}^{N_u} \frac{1}{\eta} \ln\left(1 - \eta Y_i\right)\right], \tag{4.17}$$

with $k = -\frac{1}{N_u} \sum_{i=1}^{N_u} \ln\left(1 - \eta Y_i\right)$ [25], which leads to coupled equations for the ML estimator:

$$\frac{1}{\eta_{ML}} - \left(\frac{1}{k_{ML}} + 1\right) \frac{1}{N_u} \sum_{i=1}^{N_u} \frac{Y_i}{1 + k_{ML} Y_i} = 0, \tag{4.18}$$

$$k_{ML} = \frac{1}{N_u} \sum_{i=1}^{N_u} \ln\left(1 + \eta_{ML} Y_i\right). \tag{4.19}$$

The ML estimates are obtained numerically using non-linear optimization techniques.

4.7.2 Method of Moments estimation of the GPD parameters. Method of moments (MOM) estimator [32] is a direct generalization of the Hill estimator [17], which is a form of the maximum likelihood estimator [3]. The parameters are estimated as $k_{MOM} = \frac{1}{2}\left(1 - \frac{\bar{Y}^2}{s^2}\right)$ and $\sigma_{MOM} = \frac{\bar{Y}}{2}\left(1 + \frac{\bar{Y}^2}{s^2}\right)$, where $\bar{Y}$ is the mean of the excesses and s^2 is the variance. This method depends on the relationship of the parameters k and σ to the population moments. However, the MOM estimates are not reliable for $k < -0.2$ [31]. Also, the r^{th} moment for a GP density exists only for $k < \frac{1}{2}$. Therefore, for values above $k = \frac{1}{2}$, the MOM estimator can not be obtained.

MOM is related to the ML estimator through the Hill estimator. The Hill estimator is a tail-index estimator which takes the log-spacings of extreme order statistics and uses this information to estimate k. It is

$$k_{u,n,Hill} = \frac{1}{u}\sum_{j=1}^{u} \ln X_{n-j+1,n} - \ln X_{n-u,n}, \tag{4.20}$$

where u is the threshold number and n is the n^{th} order statistic greater than u. This expression may be derived from the ML estimator formulation. With $\sigma = 1$ and Y_j the j^{th} order statistic above some threshold value Y_u, the log-likelihood of the GP density conditioned on a number of samples N_u above a threshold u is (compare Equation (4.14)) [3]

$$\ell(k|Y) = -N_u \ln k - \left(\frac{1}{k} + 1\right)\sum_{i=j}^{N_u} \ln Y_j. \tag{4.21}$$

Taking the derivative and setting equal to zero leads to

$$\hat{k} = \frac{1}{N_u}\sum_{i=j}^{N_u} \ln Y_j, \tag{4.22}$$

which is the Hill estimator for $N_u = u$. Hence, the Hill estimator is a generalization of the ML estimator. The MOM estimator is the weighted Hill estimator [17]

$$k_{MOM} = k_{u,n,Hill} + 1 - \frac{1}{2}\left(1 - \frac{k^2_{u,n,Hill}}{k_{u,n,Hill^2}}\right)^{-1}, \tag{4.23}$$

where

$$k_{u,n,Hill^2} = \frac{1}{u}\sum_{j=1}^{u} \left(\ln X_{n-j+1,n} - \ln X_{n-u,n}\right)^2. \tag{4.24}$$

The MOM estimator is generally located in a region near the maximum shown in Figure 4.9.

4.7.3 Probability Weighted Moments Estimation of the GPD Parameters.

Based on the same principles as MOM and, therefore, also a form of the ML estimator,

the Probability Weighted Moments (PWM) estimator finds k and σ from a given data set [31]. This method is a "main competitor" to the ML estimation method [16]. PWMs are generalizations of the moments of a distribution that give increasing weight to the data in the tail. The PWM estimates are $k_{PWM} = \frac{a_0}{a_0 - 2a_1} - 2$ and $\sigma_{PWM} = \frac{2a_0a_1}{a_0 - 2a_1}$, where $a_r = \frac{1}{n}\sum_{i=1}^{N_u}(1 - p_{i:N_u})^r Y_i$, $p_{i:N_u} = (i + \zeta)/(N_u + \delta)$ and ζ and δ are constants of optimization. The main limitation of this method is that PWMs for GPDs exist only for $k < 1$ [32]. Since PWMs are weighted moments-based estimators, their performance region is similar to that of MOMs.

4.7.4 ML, MOM, and PWM Estimator Performance on Simulated Data. Simulated data are generated and the methods described in the previous section are used to estimate the tail-index parameter k. Specifically of interest is the behavior of the estimator as the threshold (or "cutoff" value) u varies. The simulations are performed on a range of $0.1 - 3.2$ for the k values. Also, for $k = 0$ the GP density reduces to the exponential distribution, where the "estimation of σ is trivial" and, hence, k estimation is trivial [10]. Therefore, k values approaching zero are not considered here.

The first data set is generated using the GP density. Here 10,000 random data points are generated (with a specified k value, $\mu = 0$, and $\sigma = 1$), and each estimation method is applied to varying sizes of excess RVs (u is varied from 10 to 2000 in increments of 20). The value of k is then varied, and the estimation methods are repeated at the thresholds discussed above. The results are in Appendix B.

The next data set is generated using the $|t_\nu|$ distribution. Again, 10,000 random data points are generated (with a specified k value ($\nu = 1/k$), $\mu = 0$, and $\sigma = 1$) and each estimation method is applied to varying sizes of excess RVs (u is varied from 10 to 2000 in increments of 40). The value of k is then varied, and the estimation methods are repeated at the thresholds discussed above. The results are in Appendix C.

4.7.5 Initial Analysis. Based on the results in the appendices, the ML estimator is more accurate for threshold values incorporating a large excess sample size. For the values of u resulting in small excess sample sizes, the estimates of k tend to vary over a wide range. The estimator, however, converges to the correct value as the sample size increases. Also, the ML estimator behaves erratically for the smaller sample sizes at larger k values. For smaller k values ML tends to estimate values close to the true k value, but it tends to converge to the explicitly imposed upper bound ($k = 5$ in this case) and to a wider range of estimates away from the true k value for larger tail-index cases. As stated in [16] "ML estimates of the parameters of the generalized Pareto distribution are sensitive to the most extreme observations." The case for these most extreme observations increases as the tail-index increases (although they are fewer in number). These values tend to cause the erratic estimator behavior mentioned above.

The PWM and MOM estimators are not valid above their self-imposed k limits. However, the PWM estimator performs well in the region $0.5 < k < 1$. Also, the PWM estimator performs comparably to the ML estimator (a desirable property of the PWM, which has made it an attractive alternative to ML in the regions below its k limit). The same is true for MOM estimator in the region $0.5 > k > 0$. The main advantage of the MOM and PWM estimator is their "smooth" behavior.

4.7.6 Further Simulations and the Elemental Percentile Method. The weaknesses of the methods described in Section 4.7, based on their performance shown in the appendices, have motivated the development of other methods. The Elemental Percentile Method (EPM) [3] attempts to converge to a more robust estimate of the tail-index by combining two steps that provide information about the behavior of the data. The first step obtains many estimates from percentile and spacing information using different pairs of data points, and the second step combines these estimates to obtain a more efficient estimate.

Specifically, if $x_{i:n}$ and $x_{j:n}$ are two distinct order statistics, in the first step percentile values are equated to their corresponding CDF by $p_{i:n} = (i-a)/(n+b)$, where i is the i^{th} position, n is the total number of data points, and a is usually 1 while b is 0 [10]. These values are used to obtain spacing information $\ln(1 - x_{i:n}/\delta) = kC_i$, and $\ln(1 - x_{j:n}/\delta) = kC_j$,, where $C_i = \ln(1 - p_{i:n})$. These two expressions can be solved using the bisection method on an interval determined by an initial estimate of δ [10], where the current estimate for k is $\hat{k}_{i,j} = \ln(1 - x_{i:n}/\hat{\delta}_{i,j})/C_i$. After a value for k is calculated using all distinct pairs, in the second step the final estimate of k is determined as the median of the distribution of individual estimates of k from the distinct pairs.

One immediate limitation of this method is computational inefficiency due to the large numbers of distinct data pairs for large samples of order statistics. For this reason, in the following simulations (unlike in previous simulations) the threshold u is set to allow a maximum of 500 exceedances. This condition also allows comparison of the estimation methods for smaller threshold values of u (since, barring intrinsic limitations, the estimators tend to perform well with larger data sets). Figure 4.22 shows a plot of each estimator for $k = 0.5$, $\mu = 0$, and $\sigma = 1$.

Four tables, each showing the performance of a different estimation method, are given in Appendix D for three sets of simulations. The first simulation generates a GP density with known k values and estimates k using the methods described here for different threshold values up to $u = 500$. The second and third sets of simulations are performed for $|t_\nu|$-distributed data and F-distributed data. The data in the tables represents the difference between the estimated k parameter and the actual value for k (bias).

4.7.7 Observations. The results in Appendix D show that the EPM estimate has a high rate of variability for smaller data sets, and although it is applicable for any value of k (no constraints on k as for MOM and PWM estimates), ML outperforms EPM estimation in the range investigated here. As mentioned previously, using EPM

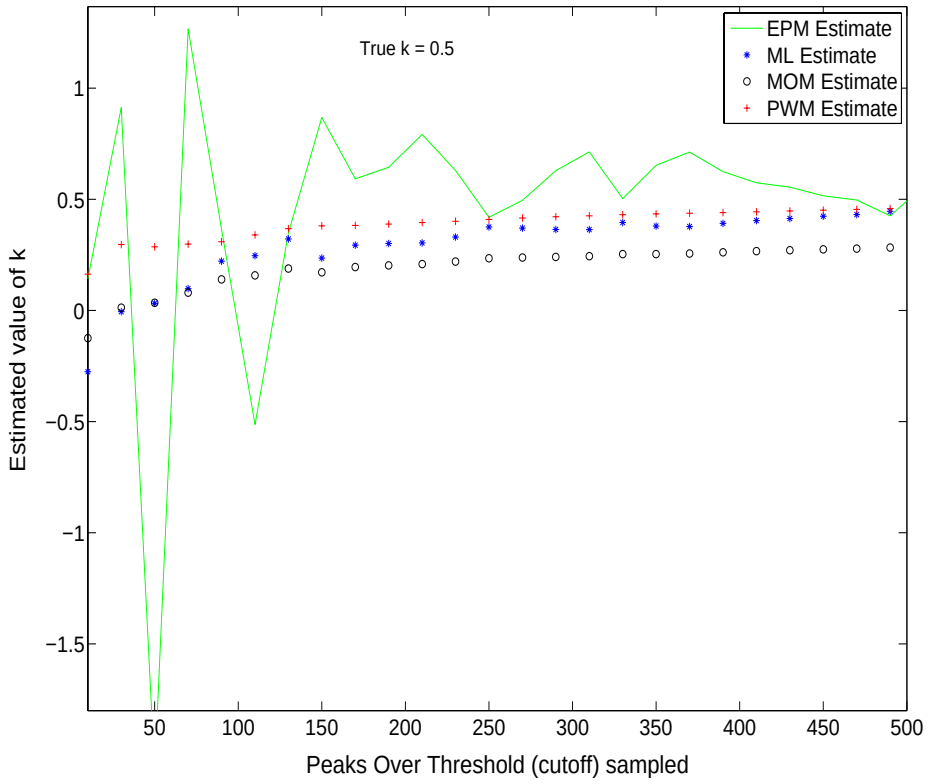


Figure 4.22: Example of different k estimation methods, where the actual value of k is 0.5. Notice the high variability in the PWM estimate, while the ML and PWM estimates are comparable. For this case, the MOM and PWM estimates exhibit the least amount of variability at different thresholds.

estimation with larger data sets poses computational problems. Sampling the distinct data pairs at different rates (not using every distinct data pair but varying intervals of pairs) might mitigate these problems, but here using different rates of distinct data pairs offers no improvement in estimation, and in most cases it yields degraded performance.

The results also show that the ML estimator outperforms the other techniques in most cases. The attractiveness of this estimator, although it yields less accurate estimates at the lower sample sizes compared to the initial simulations, is evident. Not demonstrated here is that the well-studied asymptotic properties of ML estimation provide a convenient method of confidence interval estimation. The Bayesian method

is the only method whereby full additional information (including confidence intervals) may be obtained.

As is apparent in the simulations, each estimator performs differently given smaller data thresholds. In the next section threshold selection is investigated, and sensitivity to threshold selection for each estimation technique is parameterized. Finally, guidelines on how to choose the threshold based on initial measurements that indicate in which region the tail-index may occur are developed.

4.8 Threshold Sensitivity

In this section the threshold sensitivity of the Bayesian, ML, and Hill estimators are analyzed. These estimators are selected based on their ability to perform over a wide range of k values (unlike MOM and PWM estimates).

4.8.1 Sensitivity of the Bayesian Estimator. The Bayesian estimator, as described in Section 4.5 provides a posterior pdf for k. Given a uniform prior, the maximum of the posterior pdf is the location of the ML estimate. Here, the Bayesian estimator uses a gamma prior. Sensitivity to threshold selection is monitored as a function of threshold and sensitivity to prior shape. The maximum of the posterior is the MAP estimator in this case. Figure 4.23 shows the performance of the Bayesian estimator as a function of threshold on a sample data set drawn from a GP density with $k = 1.0$, $\mu = 0$, and $\sigma = 1$.

A top-down view of the surfaces given in Figure 4.23, with only the maximum of the surface plotted, provides a two-dimensional plot of the MAP estimate of k as a function of u as shown in Figure 4.24. Notice the increased variance in the estimate at smaller threshold values. Since the second derivative of a function measures the rate of change of that function, this variance can be modeled by observing the second derivative of the estimator with respect to u. Figure 4.25 shows a profile of this phenomenon. Clearly, the second derivative is greater where the estimator shows large variance in estimation over a region of u.

Therefore, for the remainder of this analysis the second derivative of the estimator with respect to u is measured and analyzed, along with the Root Mean Squared Error (RMSE) between the estimate and the true value. The data set in the analysis consists of 1,000 data points with a threshold u between 10 and 900. In Appendix E, Tables E.1, E.2, and E.3 give results for the Bayesian estimator using three different α values for the gamma prior, and using F-distributed data with different k values. Tables E.4, E.5, and E.6 give the results for the Bayesian estimator using three different α values for the gamma prior and using GP-distributed data with different k values. Tables E.7, E.8, and E.9 give the results for the Bayesian estimator using for three different α values for the gamma prior, and using $|t_\nu|$-distributed data with different k values.

4.8.2 ML Estimator Threshold Sensitivity. From the initial analysis of GPD tail index estimator, it is evident that the ML estimator provides robust solutions for large samples from the tail of a density. Specifically, in the region $-1/2 < k < 1/2$ the ML estimator is invariant, consistent, and asymptotically normal [16]. However, because the maximization of the likelihood function requires numerical optimization, and due to extremely flat portions of the function in some cases, the estimator has limitations below certain thresholds. Also, the presence of local maxima may result in less than optimal performance and may lead to bias error.

Thus, the sensitivity of the ML estimator is examined, since in many cases only small sample sizes are available. For many small sample sizes better estimators may be obtained, as the attractive features of ML estimation are only valid for asymptotic ranges. Therefore, knowing acceptable limits for the ML estimator with small sample sizes is important.

Implementation of the ML estimator follows the approach outlined by Dargahi-Noubary and Beirlant, et al. [3, 15]. This approach requires sampling the peaks over thresholds without discarding any "outliers." In some methods, the top few samples

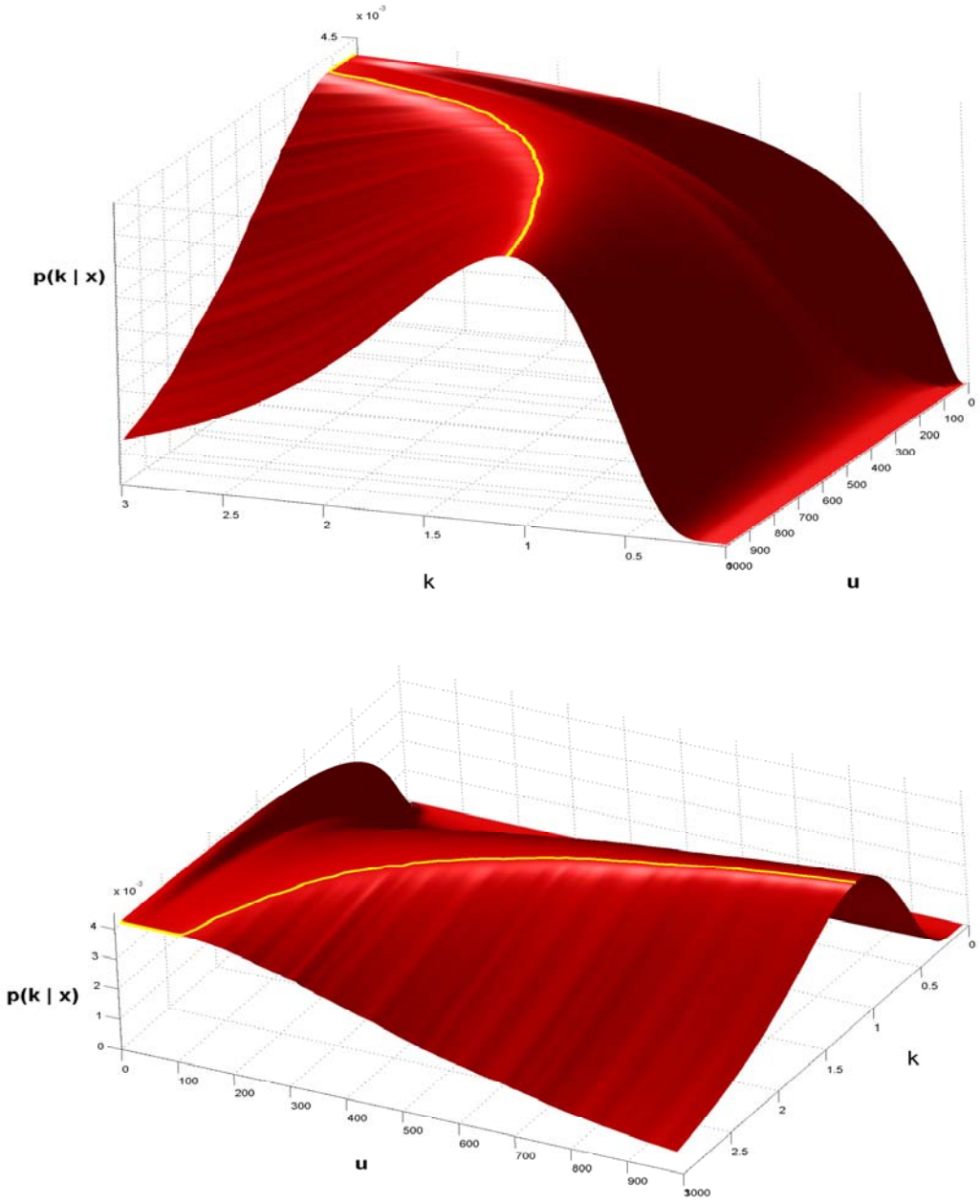


Figure 4.23: The posterior pdf surface, where the z-axis indicates the posterior value with respect to the threshold u and the values of k in the prior. The maximum of each posterior is highlighted. Notice the convergence of the peak to the true value of $k = 1.0$ as u increases. The top and bottom plots are different views.

are discarded as outliers and the remaining samples are used. However, here all of the samples are retained.

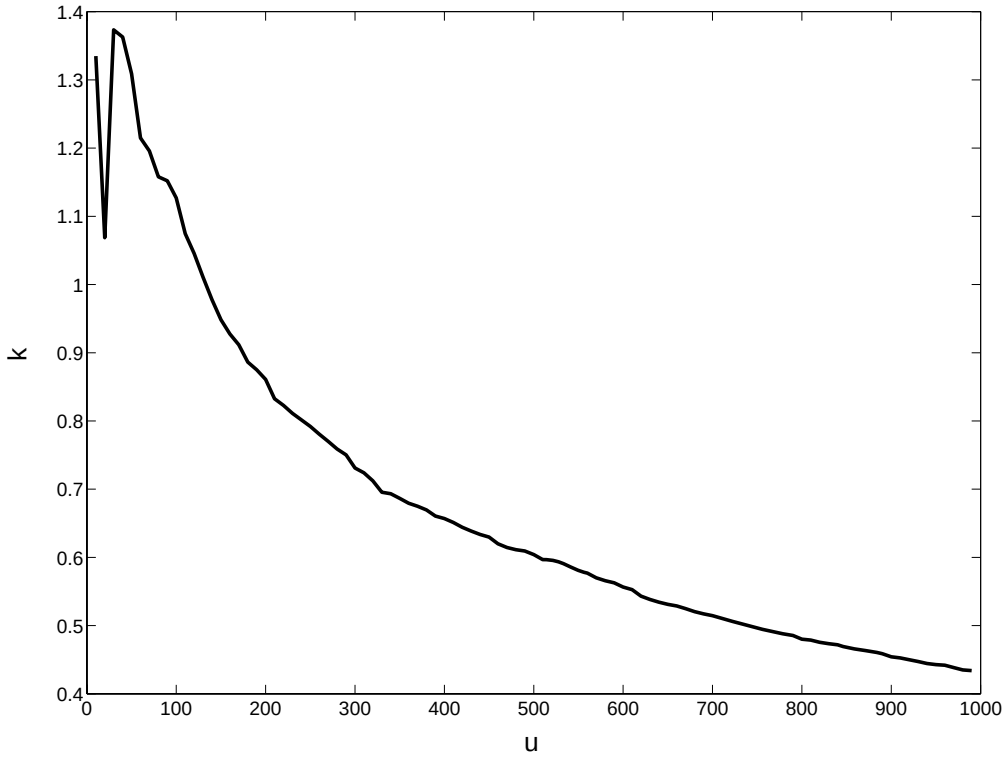


Figure 4.24: The maximum of the posterior pdf surface as a function of threshold u. Here the gamma prior is used with $\alpha = 1$ and $\beta = 1$, and thus this plot is of the MAP estimate of k as a function of u.

In the same manner as for the Bayesian estimator, the threshold sensitivity of the ML estimator is analyzed by observing the behavior of the second derivative of the estimate function with respect to u. Figure 4.26 shows the performance of the ML estimator for four different values of k. Table E.10 of Appendix E gives the results of sensitivity analysis for different regions of u.

Note that as $k \to 0$, the GP model reduces to the exponential model. Therefore, it is important to determine at what point the ML estimate should be used with the exponential model rather than the GP model. Table E.11 shows the Average RMSE of the ML estimator for k values approaching zero compared to nominal values of $0.1 < k < 1.0$. From this sensitivity analysis a k cut-off value is obtained for transitioning the model from GP to exponential.

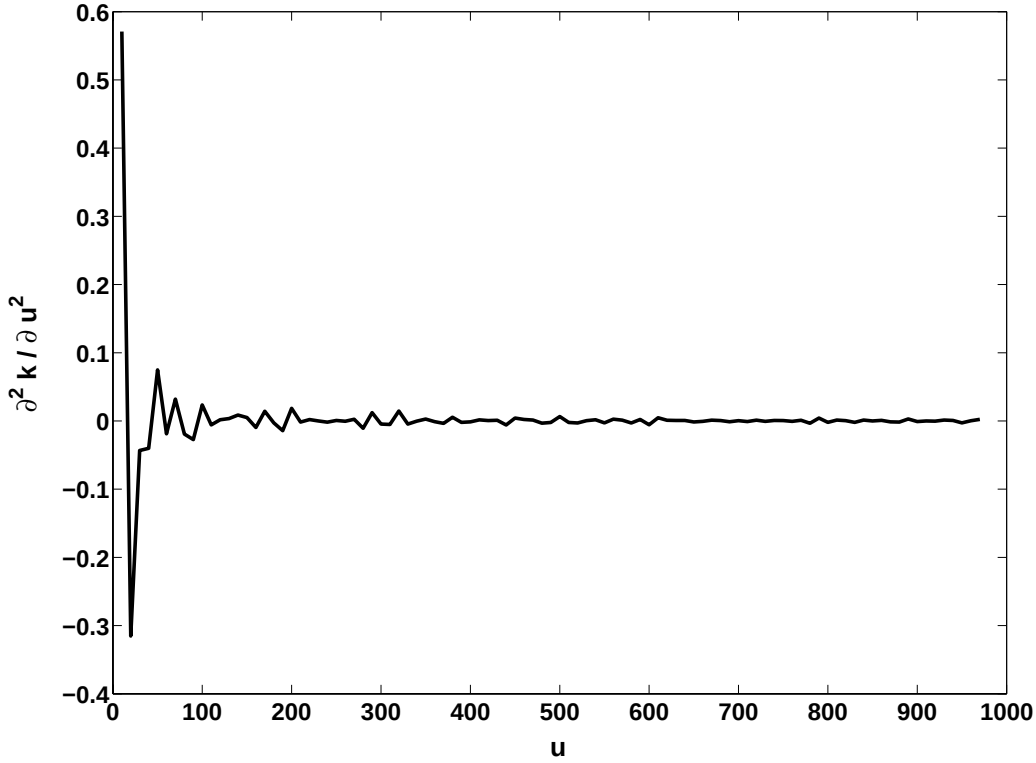


Figure 4.25: The second derivative of the estimator in Figure 4.24 as a function of threshold u. Sensitivity to threshold is evident where the second derivative is large over a region of u. For this case, choosing $u > 200$ yields decreased fluctuation in estimator performance.

4.8.3 Hill Estimator Threshold Sensitivity. The Hill Estimator results as a natural extension of studying the slope of the extreme values of an exponential plot based on the log-transformed data [3]. The least squares fit to this line results in the Hill estimator. In Section 4.7.2, the Hill estimator is also shown to result from a generalization of the likelihood function, and is, therefore, a version of maximum likelihood estimation. The Hill estimator, specified in Equation (5.3), is simple to implement, fast, and works over the entire range of k values. The last reason identifies it as a desirable estimator, along with ML and Bayesian estimators, for application in HSI data analysis. The disadvantages of the Hill estimator are its large variance and bias. Figure 4.27 demonstrates these two shortcomings.

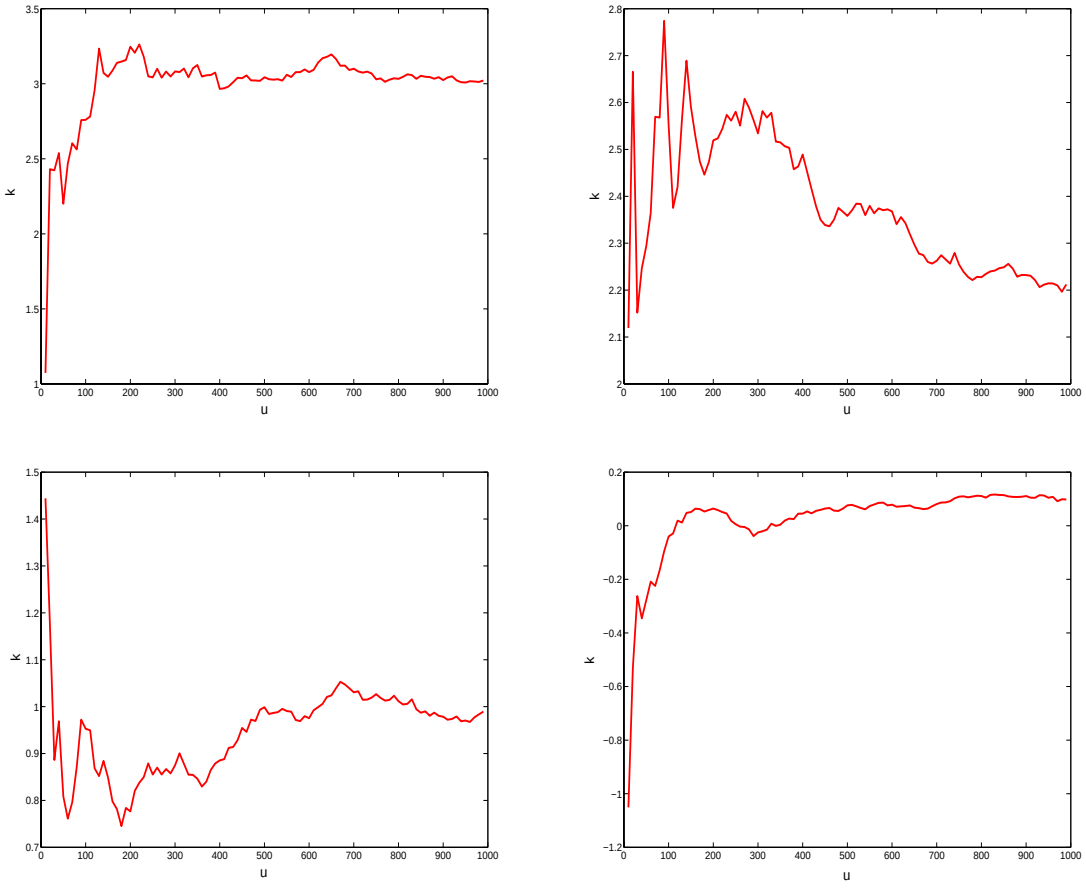


Figure 4.26: Performance of the ML estimator for $k = 3.0$ (upper left), $k = 2.0$ (upper right), $k = 1.0$ (lower left), $k = 0.1$ (lower right)

Notice from this figure that the bias in the Hill estimator increases as the threshold value increases but that the variance decreases. Also, in the smaller threshold regions the bias decreases and the variance increases, which is a classic bias-variance trade-off exhibited by many estimators. Obviously, due to large variance, the sensitivity to small threshold values is high. Therefore, there is a need to identify different forms of the Hill estimator which may mitigate the bias-variance trade-off.

4.8.4 Adaptive Threshold Selection. Modifications to the Hill estimator in an effort to mitigate the bias-variance trade-off effect and to determine optimal k values have been attempted [4, 5, 14, 26]. Most of these modified Hill estimators adaptively determine the optimal threshold by performing an initial diagnostic on a

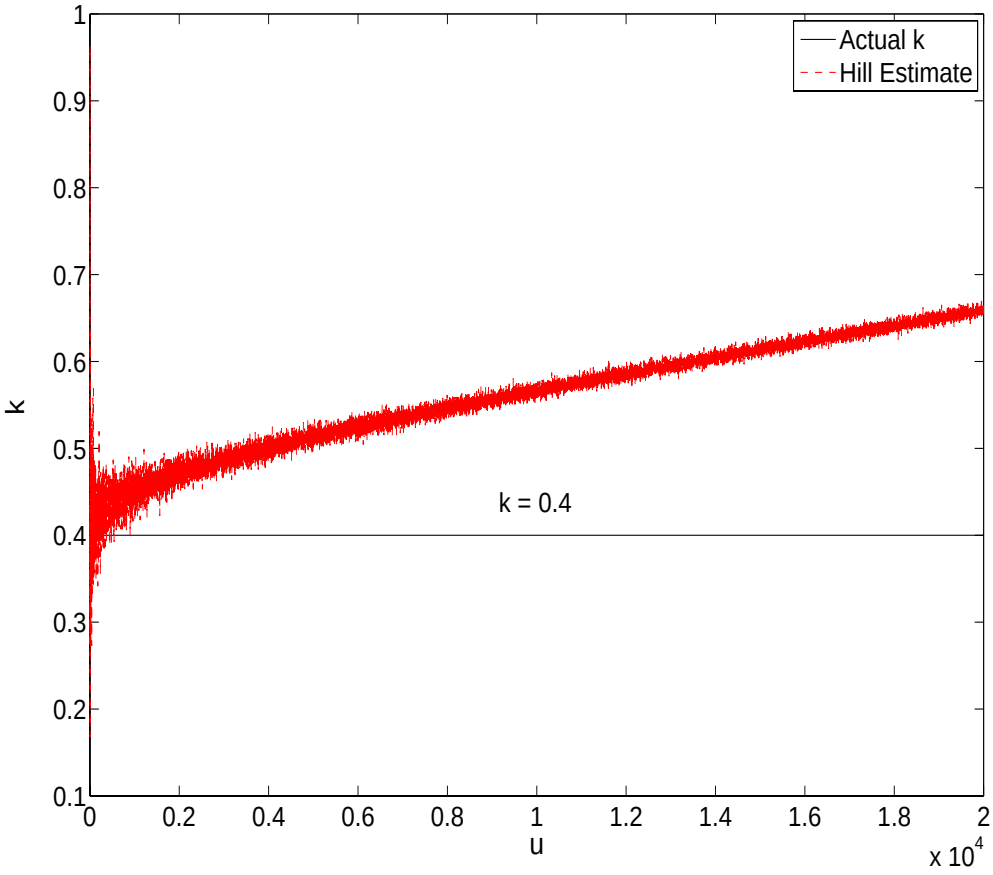


Figure 4.27: Hill estimator for k given 100,000 data points and threshold cutoffs from 1 to 20,000. The true value of k is 0.4.

sub-sample or modified sample of the data and then develop a process that selects an optimal value for k from the derived Asymptotic Mean Squared Error (AMSE) performance of the developed estimator, which often results in a weighting function on portions of the data.

The success of these adaptive methods, however, is for comparatively larger samples than considered here. For example, in [4] the adaptive methods tend to work well for samples consisting of 500 or more data points, but the authors warn that

> ...even the "optimal" Hill estimator may suffer from a substantial bias, especially for distributions with [nuisance] parameters close to 0. Therefore, it is worth considering alternative estimators for the EVI that are less

biased, such as the bias-corrected Hill estimators or unbiased estimators like the maximum likelihood or least squares estimators...

The "nuisance" parameter is a constant (difficult to estimate from a data set), which aids in providing adaptive information .

Thus, threshold sensitivity analysis for the range $10 < u < 500$ used here is not useful. The second derivative gives large values in the threshold regions, and for threshold values where adaptive methods are effective, the ML estimator is known to work well for a large range of k values. Therefore, as a conclusion from this section, the adaptive Hill estimator versions are not considered.

4.9 Threshold Sensitivity Results

Tables E.1 through E.9 in Appendix E show the performance of the Bayesian estimator for three different distributions, each with varying tail index values, and varying prior density parameters. In general the results show that as k increases the estimator RMSE and sensitivity (mean second derivative for a region of u) increase. However, as u increases, the RMSE and sensitivity decrease, as is expected. A reasonable cutoff point occurs at $u > 300$. After this cutoff the RMSE falls to acceptable values and the sensitivity becomes negligible (i.e., mean second derivative approaches zero, or equals zero in many cases).

Also, the effect of varying the α parameter in the prior is noticeable from the table results. From Figure 4.10, which displays a gamma probability density with $\beta = 1.0$ (exactly the value for the tabulated results) and varying α, the gamma prior density holds more weight in the body than in the tail at α values approaching 1.0. This phenomenon in the prior density is indicated in the tabulated results. For $k = 0.1$ the RMSE and sensitivity are less than for $k = 0.5$ and $k = 1.0$. As α increases, the RMSE and sensitivity increase for $k = 0.1$, suggesting that the Bayesian estimator is effective for $k \to 0$ tail index values with an α value approaching 1.0 when u is small, which is a result of the greater influence of the prior on the estimator for sparse data.

As k increases, the longer tails of a distribution, which result in larger extremes, cause poor performance for the estimator using smaller u values. The data set is not large enough to overcome the influence of the prior, and very little weight is placed on the large extreme values when α is such that the gamma prior density has greater body mass. Attempting to select a prior density that has the proper body and tail mass to correctly weight the influence of the larger extremes, as well as the nominal extremes, is difficult. Therefore , the use of Bayesian estimation in determining the tail index should only be used when k is suspected to be less than 1.0, where α may be adjusted such that the body and peak of the posterior density occur near the expected k value. As mentioned in Section 4.3.2, an initial guess at k can be made with the empirical quantile ratio function.

Tables E.10 and E.11 show the performance of the ML estimator with respect to different thresholds for the data. Using the $k = 0.1$, $k = 0.5$, and $k = 1.0$ tail index values mentioned in the Bayesian estimator results discussion, the ML estimator demonstrates smaller RMSE values but comparable sensitivity values. Therefore, the range for k is expanded and only the GP distribution is used in analysis. Table E.10 shows how the ML estimator performed over the expanded range for k.

Again, the region of $u > 300$ shows decreased RMSE for the ML estimator and reasonably smaller mean second derivative values. Notice the smaller RMSE values out to $k = 3.0$ compared to the RMSE values of the Bayesian estimator at $k = 1.0$. The ML estimator is not as greatly influenced by the larger extreme values. However, the influence is still present, and mitigation of this influence is addressed in the following section.

4.10 Summary and Findings

Bayesian estimation and ML estimation (which is a subset of the Bayesian process) show comparable performance and threshold sensitivity for $0.1 < k < 1.0$. However, *a priori* information drives the performance of the Bayesian estimation

method. Thus, for smaller data sets, with k increasing from zero, the ML estimator is preferred, especially at large k values.

Optimally, the least error for the ML estimator is observed in the region around $k = 0.5$ to $k = 1.0$, and this result is consistent with the literature. As k increases above 1.0, the greater extreme values begin to influence the estimation process. Many algorithms in the literature opt to exclude a certain number of the highest extreme values, thereby limiting the effect of potential "outliers."

For HSI, there is no good rule for excluding possible "outliers". Therefore, the ML estimator may suffer if the data demonstrate very large maximal extremes. One way to mitigate this effect is to initially compare log-spacings of the data for many natural scenes and scenes containing contaminations in natural scenes (which would appear as excessive deviations in log-spacings from the majority of the extremes in the tail of a distribution for the data set). This approach is similar to the initial empirical quantile ratio method introduced in Section 4.3.2 for obtaining an initial guess of k. A threshold in data log-spacing may be derived under different background conditions for an HSI data set, thus mitigating of the effect of potential outliers on the ML estimate of k. Also, the ML estimator may be modified such that the data are weighted proportional to their log-spacing information to compensate for overly large extreme values.

In summary, it is shown that a threshold not less than roughly one third of the data set leads to reasonable estimation of the tail index parameter using peaks over this threshold from a data distribution of 1,000 points. It is also shown that the ML estimator performs optimally at the thresholds for tail index values in the region $0.1 < k < 1.0$. In certain cases, with $k \to 0$ and an expertly selected prior, the Bayesian estimator performs optimally, and is desirable in that it provides all of the metrics (e.g., second and higher moments) associated with estimation given the data set. However, for small data sets the prior influence overwhelms the data information and causes inaccurate estimation, especially for large k values. The ML estimator

may be aided with *a priori* information about potential "outliers" in order to either exclude data points that adversely affect the estimator, or the ML estimator may be weighted such that these adverse effects are minimized.

Also, typical HSI data can range from a few thousand data points to over 30,000 data points per background cluster. In this case the results obtained here would be degraded, as larger data sets increase the probability of larger extreme values, which adversely affect estimator performance. Therefore, conservatively, the threshold should be increased to slightly greater than one third of the data set in order to compensate for the greater probability of adverse influence.

Practically, however, computational efficiency may suffer when one-third or more of very large data sets is used. From the tables in Appendix E, note that subsets on the order of 0.1 times data size to 0.2 times data size still result in estimates of reasonable (acceptable) deviation. Thus for larger data sets, values that tend towards 0.2 times data size should be used.

Finally, the estimator identified by this analysis for applicability to HSI MD data model estimation is the ML estimator. It is computationally efficient, not limited to specific values of k, and provides mechanisms whereby optimization, with respect to "outlier effects" may be implemented. The next chapter describes the optimized version of this estimator and its variant for use in robustly estimating GPD tail-index parameters that model HSI MD distributions.

V. Improved Tail-index Parameter Estimators for GP Models of HSI MD Data

In the previous chapter, the ML estimator for the tail-index parameter k is identified as the best method for HSI MD distribution modeling using GPDs. One limitation of the ML method is its sensitivity to largest extreme points (possible outliers) in the data. In this chapter the ML estimation method is optimized against the effects of these data points to provide a minimum error GPD fit to HSI MD distributions.

Though initially rejected in the previous chapter as a candidate due to large bias and variance effects, the Hill tail-index estimator is also investigated and modified to account for influence by largest extremes. The Hill estimator is a generalization of the ML estimator and, hence, shares some of the same properties. It is also composed of log-spacing data information, which is exploited to optimize the estimator. Initially, the Hill estimator is modified for increased robustness. The mechanics involved in mitigating possible outlier effects is explained through the Hill estimator modification process. Then, with the data space defined by the effects individual components have on the Hill estimator, the ML estimator is optimized with respect to these effects. Both the optimized ML and modified Hill estimators are applied to HSI MD data sets and are shown to improve performance.

5.1 Introduction

The ML estimator and the Hill estimator are single-point estimators resulting from an approximation to the Bayesian estimation process. The ML estimate is the maximum of the posterior probability density function estimated from the data likelihood multiplied by values from a uniform *a priori* distribution. The Hill estimator, however, is specific to estimating the tail-index parameter of the GPD

The probability density function for the GPD was given in Equation (4.9). The Hill estimator may be obtained from the ML estimator which is derived from the likelihood function associated with the GP density. With $\sigma = 1$ for the scale parameter, with $-\infty < k < \infty$ for the shape parameter (tail-index), and with X_j

the j^{th} order statistic above some threshold value X_u, the log-likelihood of the GP density conditioned on a number of samples N_u above a threshold u is [3]

$$\ell(k|X) = -N_u \ln k - \left(\frac{1}{k} + 1\right) \sum_{i=j}^{N_u} \ln X_j. \tag{5.1}$$

Setting the derivative to zero yields

$$\hat{k} = \frac{1}{N_u} \sum_{i=j}^{N_u} \ln X_j. \tag{5.2}$$

The Hill estimator is a tail-index estimator which takes the log-spacings of extreme order statistics and uses this information to estimate k. It is [3]

$$k_{u,n}^{Hill} = \frac{1}{u} \sum_{j=1}^{u} \ln X_{n-j+1,n} - \ln X_{n-u,n}, \tag{5.3}$$

where u is a cutoff threshold, n is the sample size, and j is the j^{th} order statistic greater than u. Setting $X_j = X_{n-j+1,n}/X_{n-u,n}$ and $N_u = u$ results in the same expression as Equation (5.2). Hence, the Hill estimator is a generalization of the ML estimator.

5.2 Hill Estimator Improvement

The Hill estimator suffers from large variance for small sample sizes and tends to have a bias as sample size increases, particularly for data that includes points which deviate from the GPD model. Many adaptive forms of the Hill estimator have been developed to overcome problems associated with this bias-variance tradeoff [4, 26]. These adaptive methods involve sub-sampling an initial sample of the data set to obtain a statistic which determines whether to increase or decrease the cutoff associated with the original sample to obtain a minimum mean-squared error estimate of the tail-index (thereby minimizing bias-variance effects). However, many iterations

of sub-sampling may impede the computational efficiency of this method for larger data sets.

Also, for HSI MD data sets a variation in the pattern of the tail samples occurs not only beyond (or prior to) a single specified cutoff, where samples beyond (or prior to) the cutoff tend to provide a minimum mean-squared error framework for an estimator model. But sub-samples of MD points in separate regions of the tail beyond a single specified cutoff (here, a single cutoff threshold is specified based on an analysis of estimator sensitivity to threshold selection (see previous Chapter)), contribute to degrading mean-squared error values. Thus, the optimization process described here uses a single unadjusted cutoff threshold selection, to obtain an initial Hill estimate for k.

Then, based on the least-squares line fit to the samples and using the Hill estimate as the slope for the line, certain points in the sample are censored (based on their deviation from the majority of the sample set) in a second pass of the Hill estimation process. The censored data points are not only the most extreme points, but also data that lie within the body of the sample which may contribute to variance in the estimate. This process results in an improved minimum mean-squared error fit to the data, and it is explained in more detail in the following.

The Hill estimator results from the slope of the least-squares line fit to the largest values of an exponential quantile plot based on log-transformed data [3]. Figure 5.1 shows the upper region of 10,000 data points, randomly selected from an exponential distribution with mean = 1,000, with the least-squares fit line. It is known that this region follows a GPD, i.e., the slope of the linear portion of the larger data values is equal to the tail-index parameter of the GPD function [3].

Analytically (and from Figure 5.1) as $n \to \infty$, $\ln X_{n-j+1,n} \sim k \ln((n+1)/j)$. This relationship yields

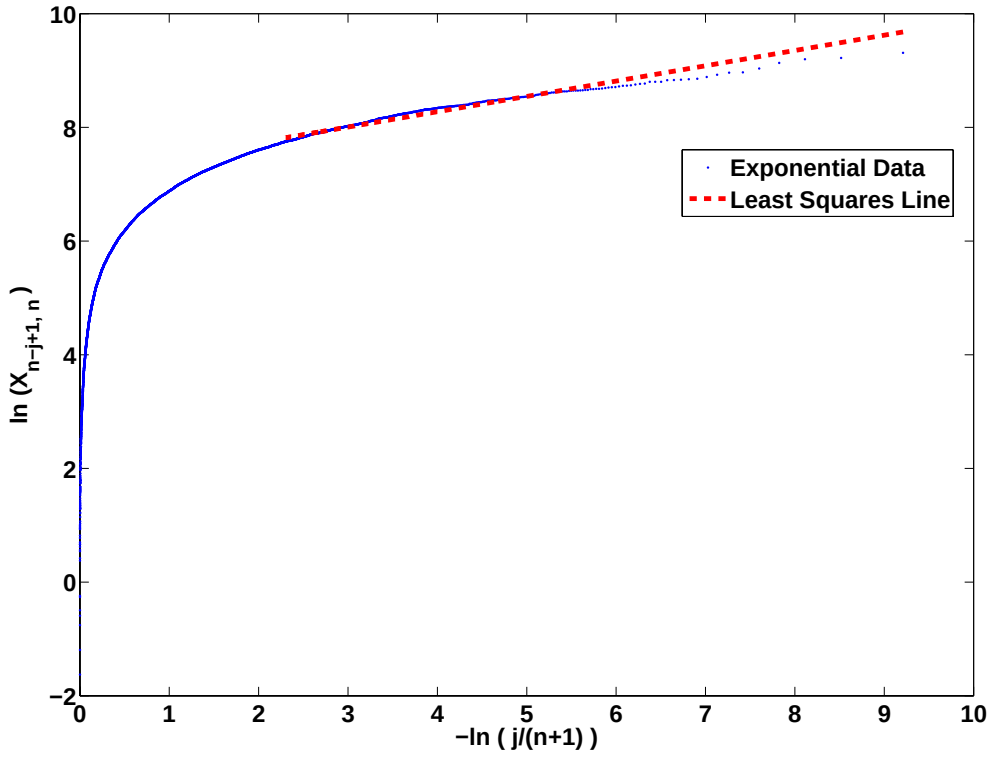


Figure 5.1: An exponential quantile plot based on the log-transformed data (see text). The line fit to the linear portion of the plot is the equation from which the Hill estimator arises. Examining the fit of this line, an optimized Hill estimator may be derived. For example, the bias in MSE is caused by regions of the plot that deviate from the linear region (i.e., the GPD model).

$$y = \ln X_{n-u,n} + k\left(-\ln\frac{j}{n+1} - \ln\frac{n+1}{u+1}\right) \tag{5.4}$$

$$= \ln X_{n-u,n} + k\left(\ln\frac{u+1}{j}\right). \tag{5.5}$$

Hence, forcing the data points toward greater linearity above a set cutoff u results in a more robust estimate of k. For example, a quantile plot of log-transformed data from two F-distributed data sets is shown in Figure 5.2. Figure 5.3 shows the data points optimized to generate the least-squares line fit for minimizing MSE. The MSE for the original data set is 0.12, while the optimized data set MSE is 0.08.

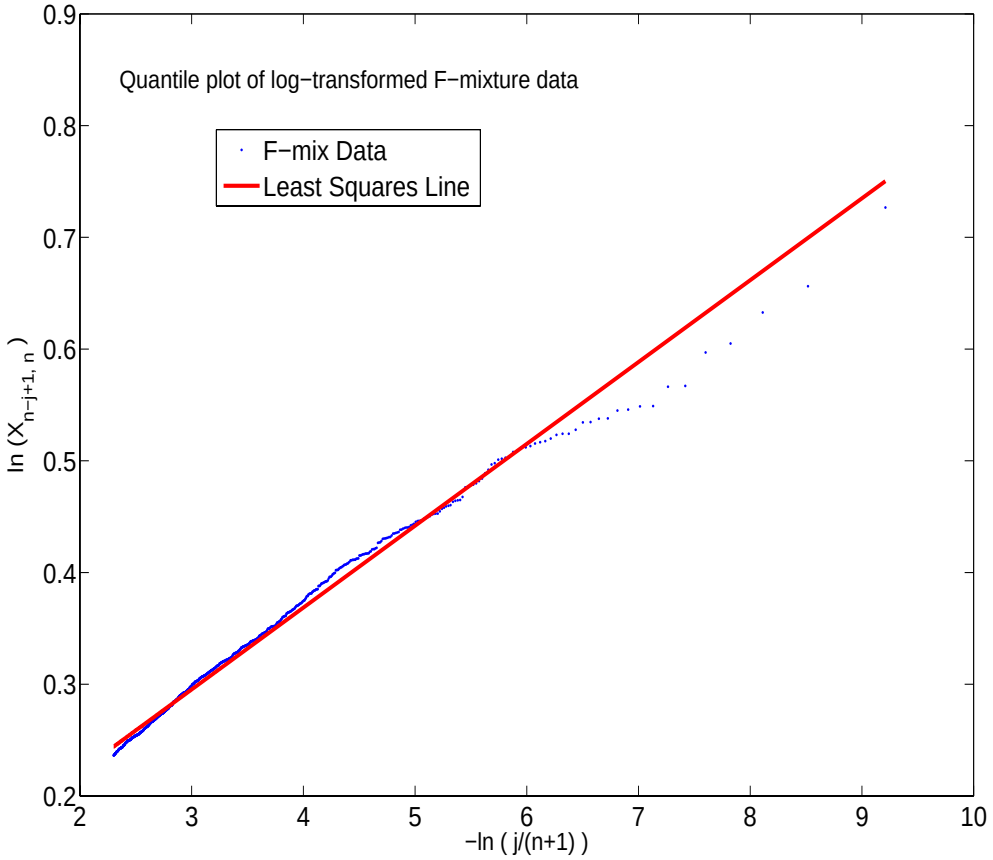


Figure 5.2: A quantile plot based on the log-transformed data from a mixture of two F-distributions with parameters: $\nu_1 = 155$ and $\nu_2 = 30$ for the first F-distribution, $\nu_1 = 155$ and $\nu_2 = 500$ for the second F-distribution, and mixing ratios of 20% and 80%. Initially, 10,000 data points are created; then the 1,000 largest values are selected for analysis. Notice the deviation from the least-squares line.

The optimization process described above is applied to HSI MD data, and an exceedance plot is generated to display the improvement in estimation of k. Figure 5.4 shows the quantile plot of log-transformed HSI MD data from a subset of data from a cluster of vegetation from Figure 3.7. Figure 5.5 shows the same plot optimized for a better line fit to the data in that the data points that deviate from the majority of the sample population are removed. Figure 5.6 demonstrates the improvement in fitting a GPD to the MD data with the k estimated by this two-step process.

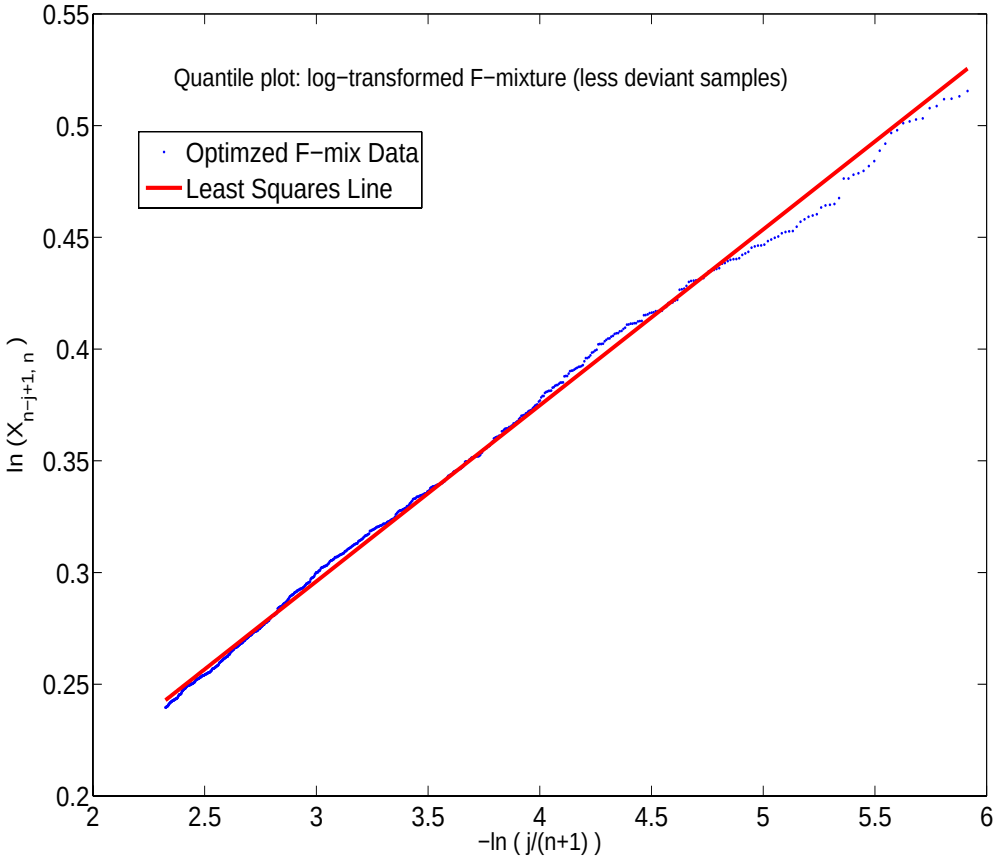


Figure 5.3: A quantile plot based on the log-transformed data from a mixture of two F-distributions with parameters: $\nu_1 = 155$ and $\nu_2 = 30$ for the first F-distribution, $\nu_1 = 155$ and $\nu_2 = 500$ for the second F-distribution, and mixing ratios of 20% and 80% and optimized to censor data points that deviate above a given distance from the majority of points. Notice the better fit to the data.

5.3 ML Estimator Improvement

The ML estimator is not limited by bias effects at large sample sizes, as observed for the Hill estimator. The performance of the ML estimator is governed by sample size inadequacy; smaller sample size results in greater bias and variance, larger sample sizes decrease the bias and variance. However, for very large data sets, as is common in HSI, when using a cutoff threshold u such that the sample contains a significant proportion of largest extreme values, the most extreme points cause the estimator to perform sub-optimally.

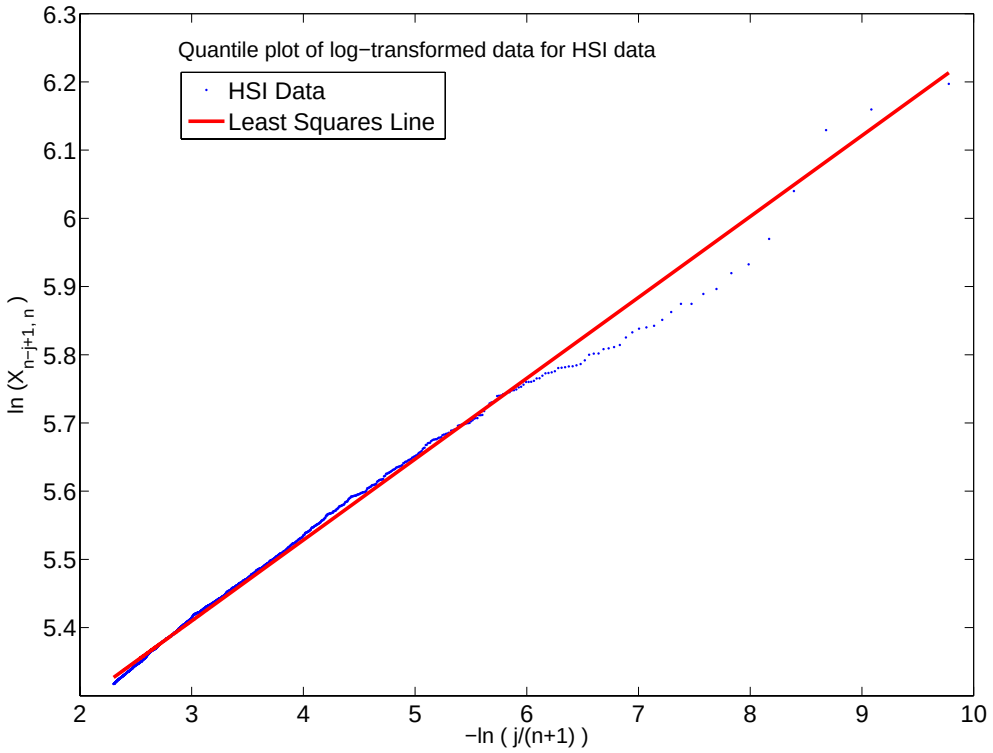


Figure 5.4: A quantile plot based on log-transformed HSI data from from a subset of data from a cluster of vegetation from Figure 3.7. The cutoff u is set such that points in the largest 10^{th} percentile are retained.

For example, GP distributed data with k approaching 0 have shorter tails. Data sets containing, 1,000, 10,000, or 100,000 data points governed by such GPDs contain few grossly extreme values. However, as k increases above 0.5, the greater amount of data in the tail increases the probability for larger data sets (the 10,000 and 100,000, as opposed to the 1,000) yielding many overly extreme data points. The large numbers of most extreme data points for larger data sets create poor ML estimator performance, and in some cases the extreme points are regarded as potential outliers [16].

HSI MD data sets tend to range in size from roughly 10,000 to 30,000. As mentioned, when k is relatively small the effects of the most extreme points on the ML estimator are negligible. For larger k the greater probability of these effects must

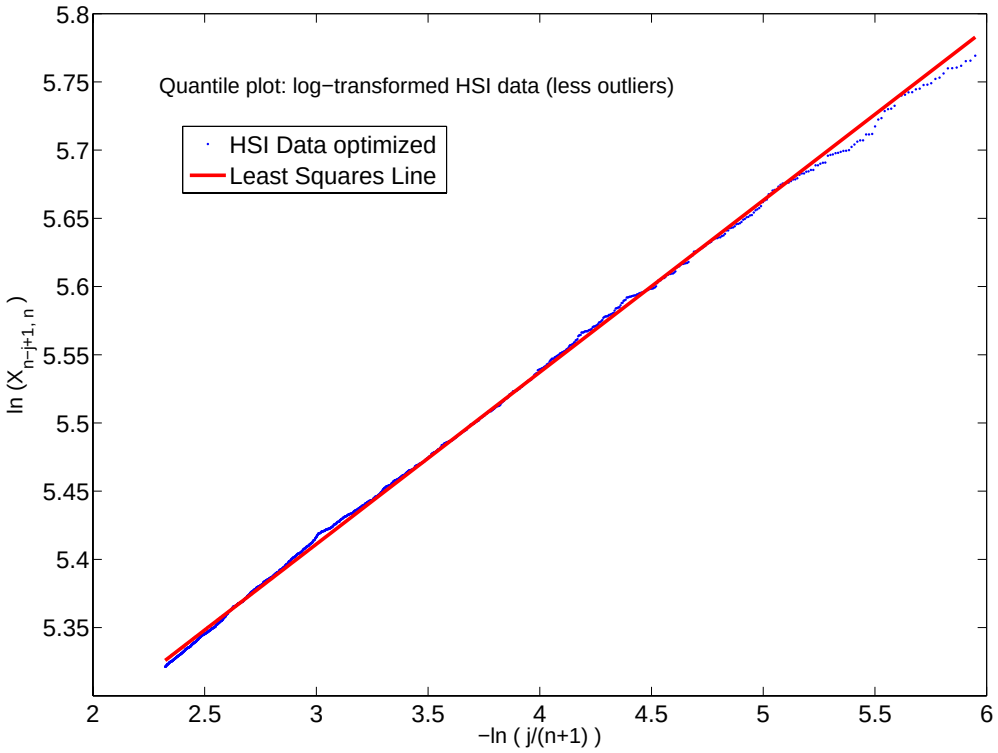


Figure 5.5: A quantile plot based on the log-transformed HSI data from Figure 3.7 optimized to censor the effect of deviations from the majority of the sample. Notice the better fit to the data.

be taken into account and mitigated. An approach for accomplishing this result is described next.

Given a data set suspected of heavy-tailed behavior (as is the case with HSI MD data), an initial estimate of the location of k for the GP model of the tail is needed. Using a general idea of the value of k, a secondary step compensates for the effect of the largest extremes if k is relatively large (if k is comparatively small the secondary step may be omitted).

An initial value for k may be obtained by using EQRF, developed in Chapter IV. This method essentially compares the distances between two regions in the tail of a distribution to the distances of quantiles in the body of the data. Once the region of k is known from the EQRF, a decision is made on whether two proceed with a

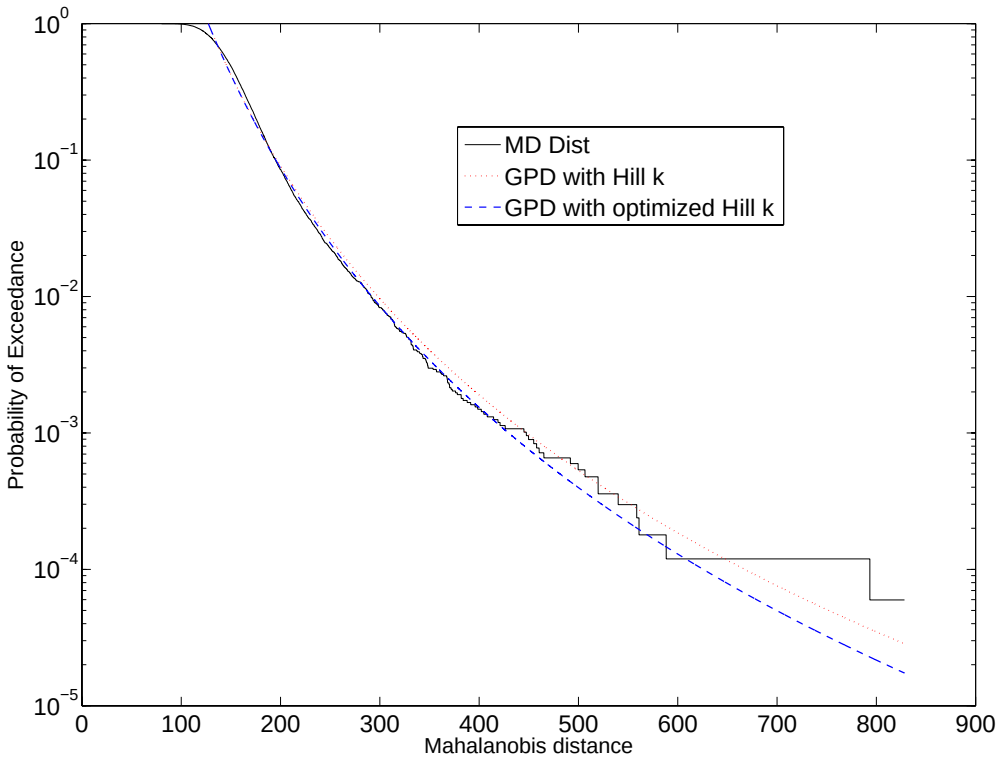


Figure 5.6: An exceedance plot of the HSI MD data from Figure 3.7 (solid line), the initial GPD fit to the data with the Hill estimate for k (dotted line), and the second-pass improved Hill estimated k GPD fit (dashed line). The MSE for the initial fit is 4.69E-04 and the MSE for the optimized fit is 4.58E-04.

second step to optimize the ML estimate of the k parameter. If the k value suggested by EQRF is relatively small, it is unlikely that the largest extremes will corrupt the performance of the estimator. However, for large enough k values the second step is taken for an optimal estimate of k.

This second step requires an understanding of how EQRF suggests and ML estimates the parameter k. EQRF tends to provide a suggested k value lower than actual because the slope of the EQRF line is based on $n \to \infty$ data points. However, the quantiles used in determining the discriminating metric are extracted from finite data sets, and therefore for progressively smaller data sets the suggested k value is lower than actual.

The ML estimator tends to estimate k values slightly higher than the actual value due to the effect of the largest outliers in data sets of sizes comparable to HSI cluster MDs. This overestimation is especially the case when the actual k value is relatively large and the data set is large. The largest extremes for data sets with large k values, tend to extend the tail of the exceedance plot of the MDs, causing the fitted model to have a heavier tail than appropriate for the majority of the data.

Given a substantially large k value suggested by EQRF, the second step requires finding a value between the EQRF-suggested k and the ML-estimated k. Since a lower limit (EQRF k) and an upper limit (ML k) are known, candidate k values between the two limits are examined to find the one that yields minimum MSE. This process is similar to the iterative steps taken in the adaptive Hill estimation processes.

HSI clusters of vegetation tend to yield very small EQRF suggested k values. Likewise, the corresponding ML estimated k values tended to be close to the EQRF values. However, a scale for selecting the midpoint to minimize EQRF is still feasible. A scale for any range of EQRF and ML k value differences may be developed once the general range of tail-index values is determined.

Although the ML estimates are close to the EQRF suggested values for the vegetation cluster, they are still affected by the largest extreme values, resulting in an estimate of k larger than the actual value as indicated in Figure 5.7. Here a typical HSI MD data set from a vegetation cluster is modeled by a GPD with the EQRF suggested k value, a GPD with the ML estimated k value, and a GPD with the optimal midpoint value. Notice that the optimal midpoint value provides the best fit.

5.4 Improved Estimators Applied to HSI Data

In this section two HSI data sets are analyzed with the improved Hill and ML estimators. The HSI data are from an AVIRIS Ft AP Hill, VA [24] scene, and the data sets are from two clusters shown in Figures 5.8 and 5.9.

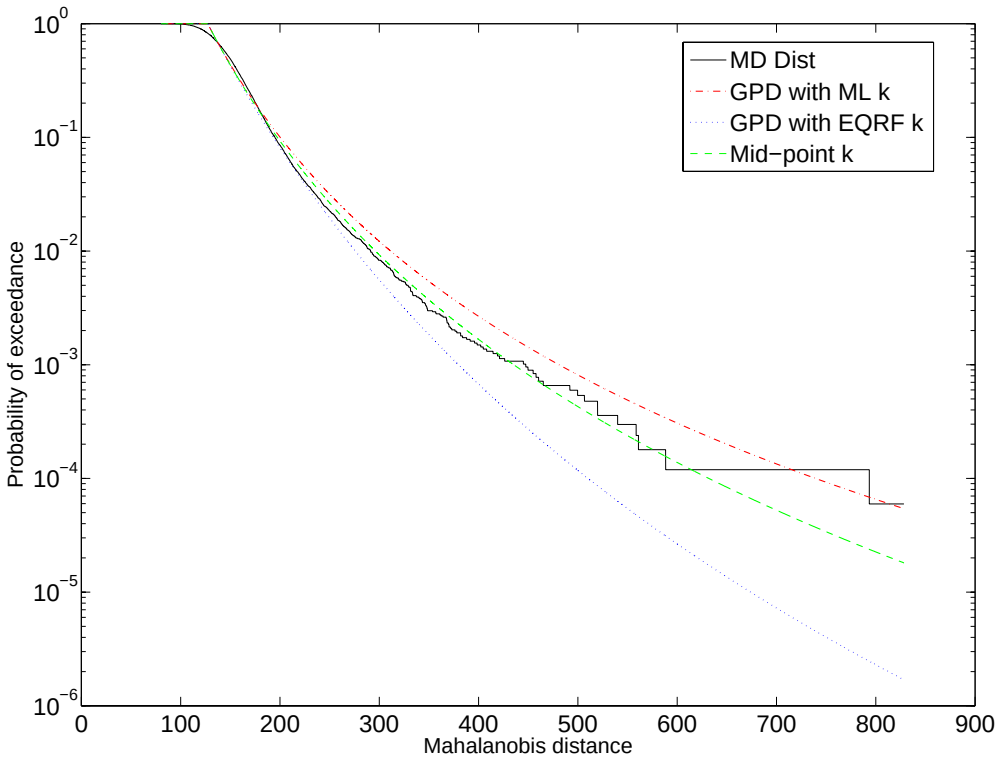


Figure 5.7: An exceedance plot of HSI MD data from Figure 3.7 (solid line), the GPD fit to the data with an ML estimated k value (dash-dot line), a GPD with an EQRF suggested k value (dottted line), and a GPD with a k value between the EQRF and ML values that provides the minimum mean-squared error fit to the data. In comparison, the ML GPD fit MSE is 0.0016, the EQRF GPD fit MSE is 0.0019, and the optimal k GPD fit MSE is 0.0015.

Both cluster data sets are reduced to MD distributions and the distributions are initially analyzed for heavy-tail behavior. The EQRF score for cluster 1 suggests the region $k = 0.11$ and $k = 0.09$ for cluster 2. Therefore, the MD distributions are identified as exhibiting heavy-tail behavior.

In applying the improved Hill estimator first to a subset of data above a threshold $u = 1000$, the first step is an initial-pass Hill estimate of the k value, which results in a value for the slope of a least squares fit line through the subset data. For example, Figure 5.10 shows the least squares line fit through this data subset. Next, certain

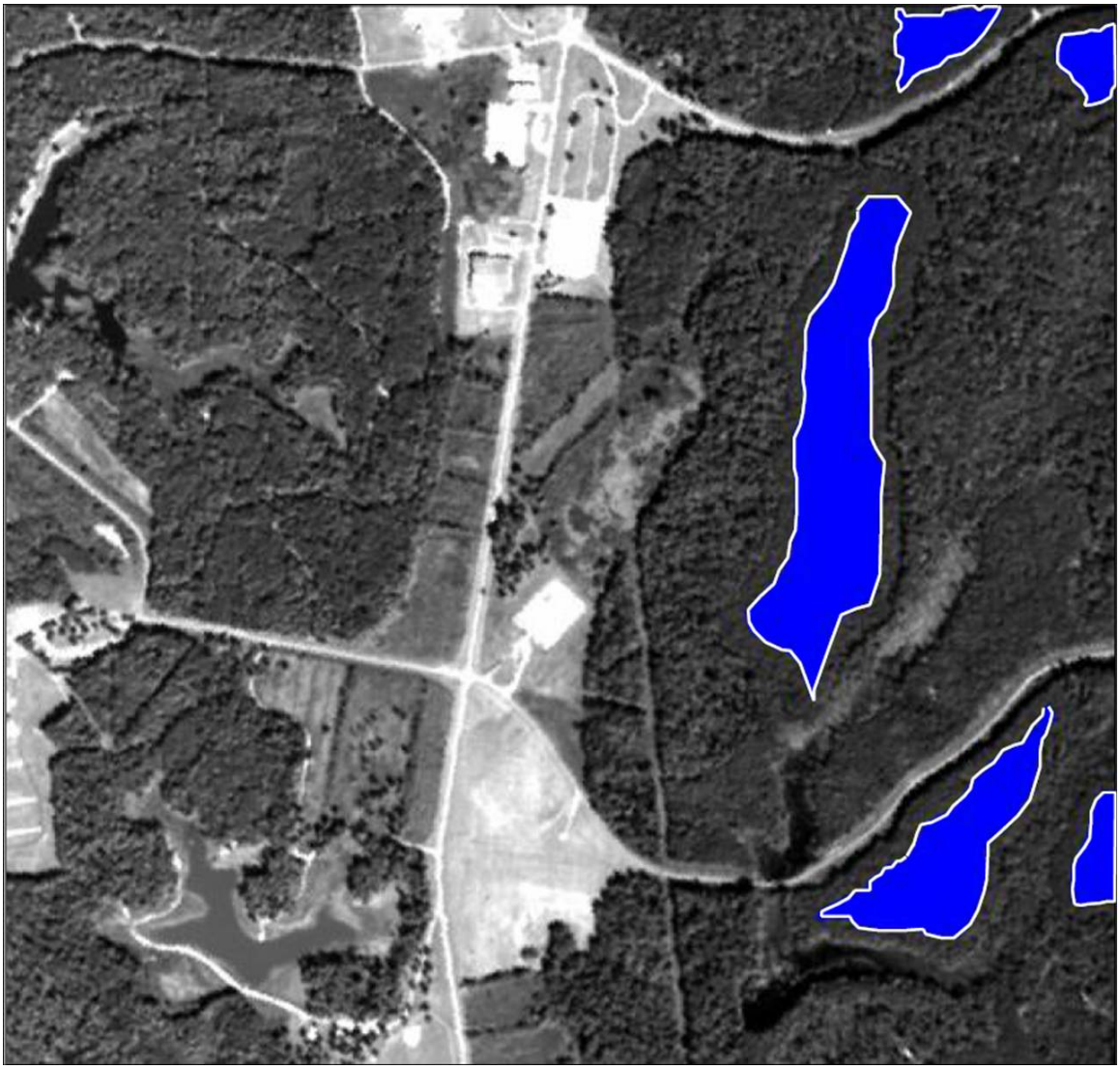


Figure 5.8: A cluster of Loblolly pine trees containing 14,257 pixels.

data points that deviate from the line fit are censored for a second-pass Hill estimate. The data are censored based on their euclidean distance from the fit line.

Using a one-norm measure for the bias and a two-norm measure for variance of the difference between the data and the fitting model, and based on results from simulations for 32 vegetative HSI data clusters using 155 bands (6-30, 34-76, 78-95, 97-101, 118-149, 179-210) out of an available 224 bands, the censoring criteria are: *a*) points with a distance greater than $\mu_{bias} + \sqrt{\sigma_{variance}}$ are excluded for MD distributions with $\sqrt{\sigma_{MD}} < 0.1D$, *b*) $2 \cdot \mu_{bias} + \sqrt{\sigma_{variance}}$ for MD distributions with

Figure 5.9: A cluster of deciduous forest containing 11,557 pixels.

$0.1D < \sqrt{\sigma_{MD}} < 0.2D$, and, rarely, c) $3 \cdot \mu_{bias} + \sqrt{\sigma_{variance}}$ for MD distributions with $\sqrt{\sigma_{MD}} > 0.2D$, where μ_{bias} is the mean one-norm distance of each point in the data subset to the fit line, σ_{MD} is the corresponding standard deviation in the distances, $\sigma_{variance}$ is the standard deviation in the two-norm distances, and D is the dimensionality of the data (in this case $D = 155$). Figure 5.11 shows the data points remaining and a least squares line fit with a slope value from the second-pass Hill estimate of k.

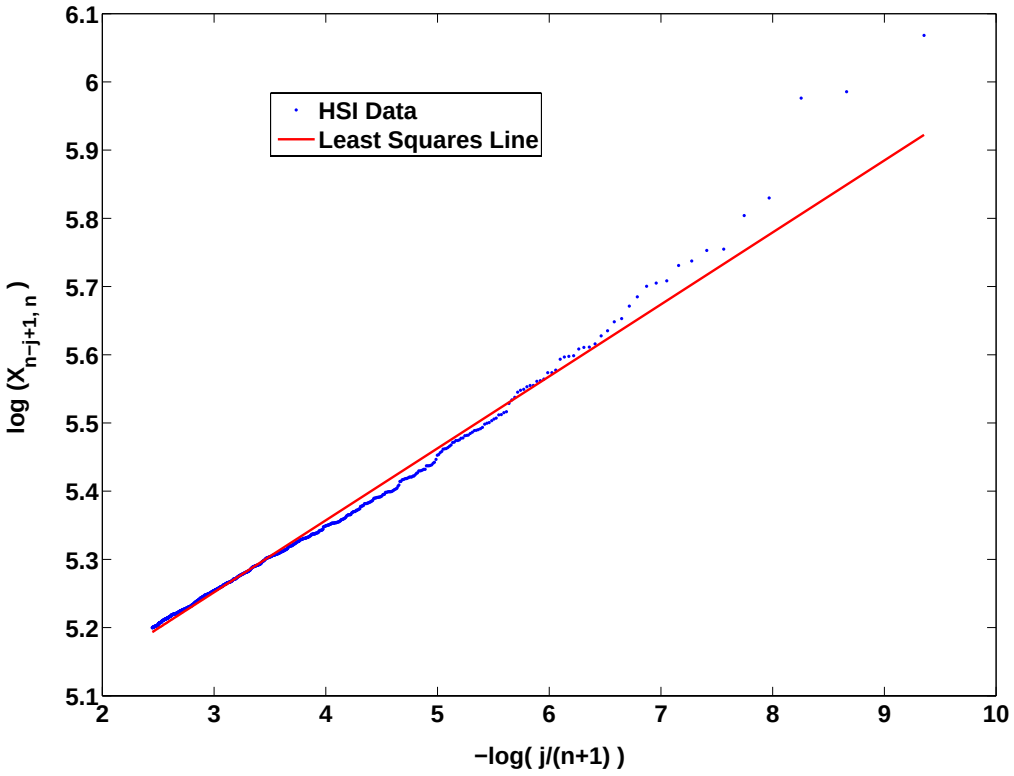


Figure 5.10: A quantile plot based on log-transformed HSI MD data from the cluster in Figure 5.8. The cutoff u is set such that the largest 1,000 data points are analyzed. Notice that the largest extreme values tend to deviate greatly from the fit line.

Cluster 2 is analyzed in the same fashion. The estimator produces an optimal k value (for $u = 1,000$) in the minimal mean squared error sense. The GPD fit to these data are shown in Figures 5.12 and 5.13. For cluster 1 $k_{Hill} = 0.12$ with a MSE for the initial fit of 2.1E-03. The MSE for the improved fit with $k_{opt} = 0.09$ is 0.7E-03. For cluster 2 $k_{Hill} = 0.1$ with a MSE for the initial fit is 3.5E-02. The MSE for the improved fit with $k_{opt} = 0.08$ is 3.3E-02. The process for the improved Hill method is outlined in Figure 5.14.

The improved two-pass ML tail-index estimator is applied to the same clusters. The analysis for cluster 1 is summarized in Figure 5.16 with the exceedance plots of the GPDs fit to the MD data with k from the ML and improved ML approach. The

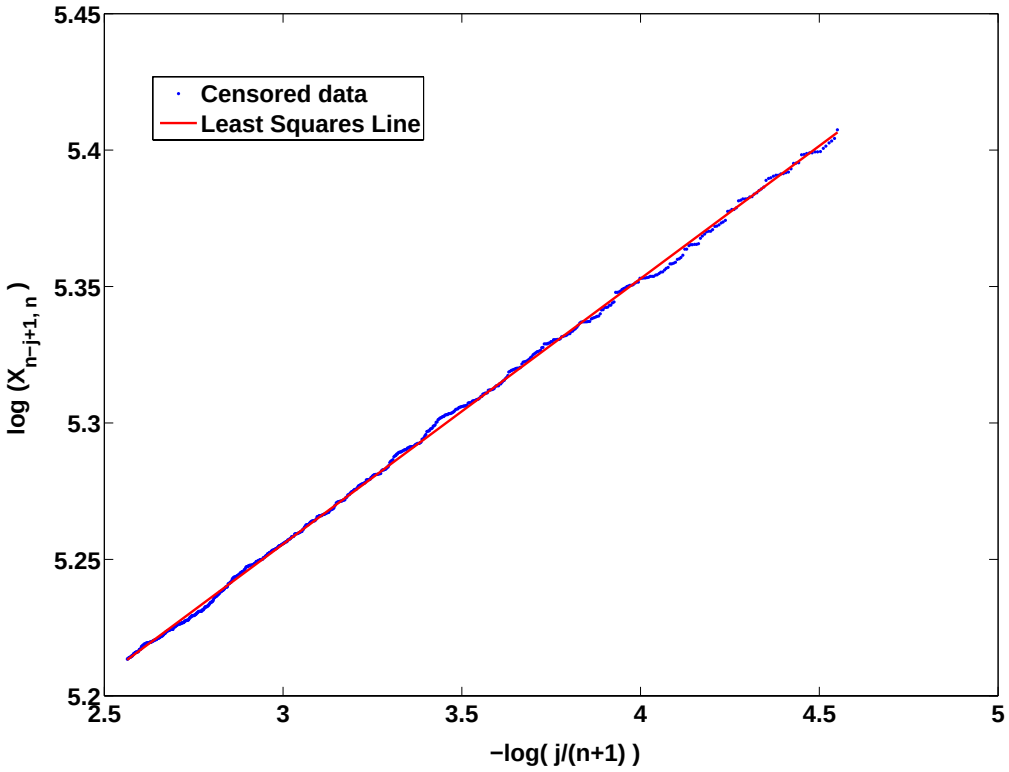


Figure 5.11: A quantile plot based on log-transformed HSI MD data from the cluster in Figure 5.8 optimized by censoring the effect of deviations from the majority of the data

analysis for cluster 2 is summarized in Figure 5.17. The process for the improved ML estimator is shown in Figure 5.15.

5.5 Results from the Improved Estimators Applied to HSI Data

Results show that the two-pass improved Hill and ML estimators provide a better fit to the entire MD distribution in the minimum mean squared error sense. The results are obtained by analyzing only the extreme data points ($u = 1,000$). Of interest is the sensitivity of estimators to the threshold u. One of the main drawbacks of many tail-index estimators is that they are highly sensitive to varying threshold values. The performance of the two-pass optimized Hill estimator is shown in Figure 5.18 for cluster 1 and Figure 5.19 for cluster 2 compared to the classic Hill estimator.

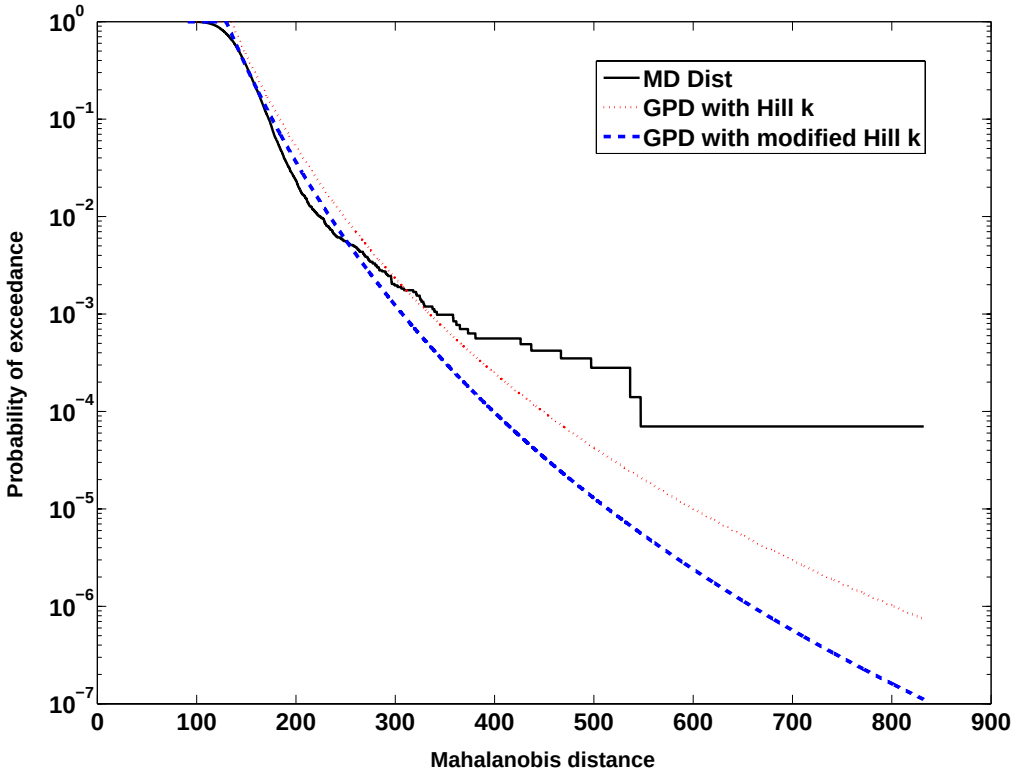


Figure 5.12: An exceedance plot of the HSI MD data (solid lines) from the cluster in Figure 5.8, the initial GPD fit to the data with the Hill estimate ($u = 1,000$) for k (thin dotted line), and the second-pass improved Hill estimated k GPD fit (thick dotted line). The MSE for the initial fit is 2.1E-03 and the MSE for the optimized fit is 0.7E-03.

The performance of the optimized ML estimator is shown in Figure 5.20 for cluster 1 and in Figure 5.21 for cluster 2.

The difference between the maximum k value and the minimum for the classic Hill estimator is 0.13 for cluster 1 and 0.04 for cluster 2. For the optimized estimator, these values are 0.11 and 0.035, respectively. Comparatively, the improved estimator is slightly less sensitive to threshold variance.

Notice also the large fluctuation in the estimator from very small thresholds to gradually increasing thresholds. This result is due to the fact that at smaller thresholds the ends of the tail are modeled, which is where (for HSI data) anomalous data appear. These data points exhibit the largest MD values, as they are most

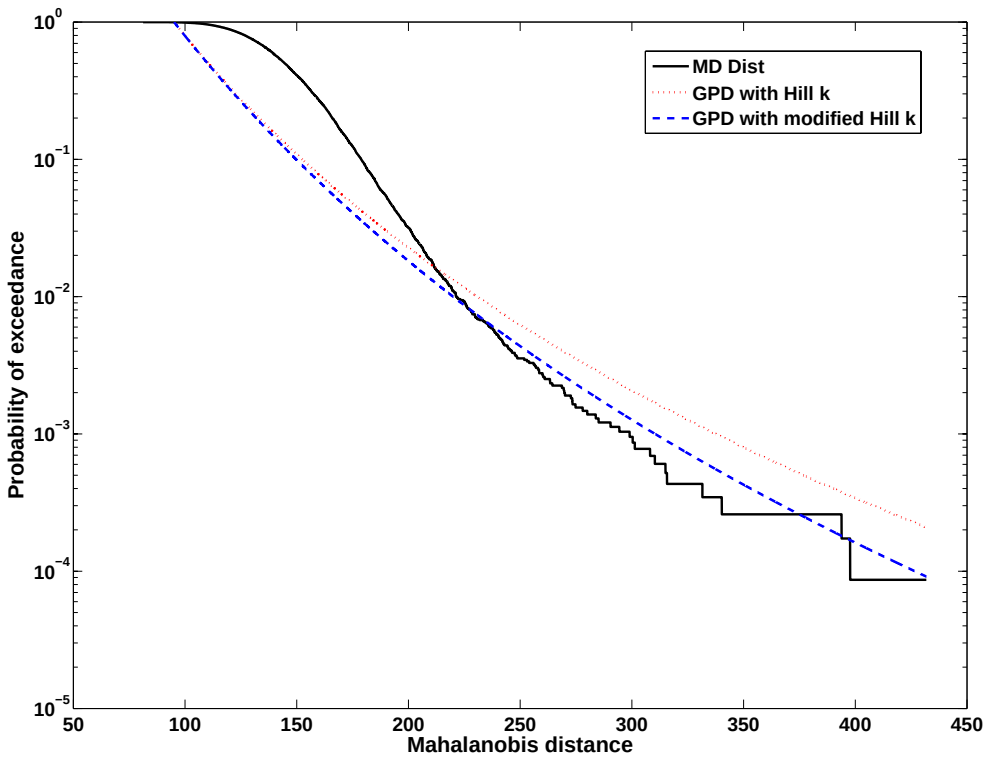


Figure 5.13: An exceedance plot of the HSI MD data (solid lines) from the cluster in Figure 5.9, the initial GPD fit to the data with the Hill estimate ($u = 1,000$) for k (thin dotted line), and the second-pass improved Hill estimated k GPD fit (thick dotted line). The MSE for the initial fit is 3.5E-02 and the MSE for the optimized fit is 3.3E-02.

unlike the rest of the cluster pixels. In this region data exploitation routines begin to discriminate between consistency in a cluster and outliers/anomalies. Therefore, for the purposes of fitting a model which correctly describes the behavior of the majority of the cluster data, abrupt fluctuations may be ignored and a rule may be developed to select thresholds above this region.

The threshold plots for the ML estimator and optimized ML estimator are shown in Figure 5.20 and Figure 5.21. The plots confirm the statement that the ML estimator improves in performance as the number of data points in the subset increases. Also, disparity between the ML and improved ML estimators decreases as the subset size increases and eventually the two estimators perform alike. This result

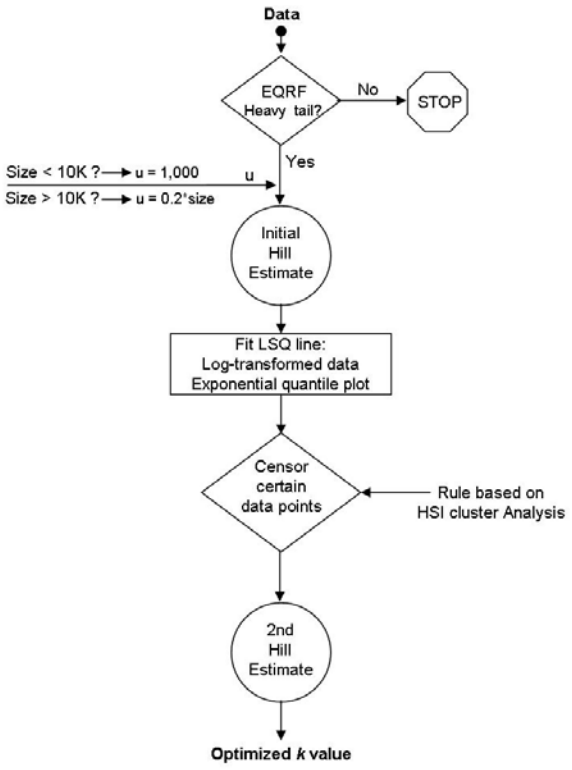


Figure 5.14: A flowchart for the process for Hill estimator optimization. The external inputs result from threshold analysis (selecting u) and cluster MD distribution analysis (determining bias-increasing data points).

is also illustrated in the plots of MSE for ML and improved ML shown in Figure 5.22 for cluster 1 and Figure 5.23 for cluster 2.

5.6 Summary

The improved ML estimator (optimized in the sense of minimizing MSE) consistently outperforms the traditional ML estimator as shown for the two examples here. In repeated application to different HSI clusters similar performance is observed. Also, as mentioned above, the ML and improved ML estimates converge in performance as the size of the data subset increases.

The two-pass improved Hill estimator provides a fit with a smaller MSE than the Hill estimator on the two HSI vegetation clusters. The two methods are comparable in threshold sensitivity, with the improved estimator showing slightly better

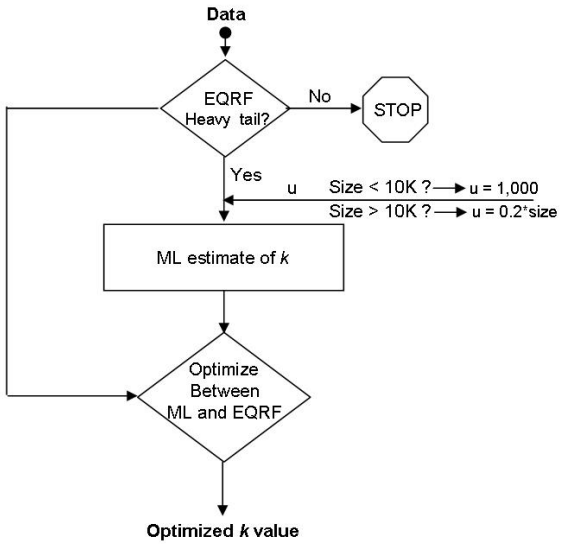


Figure 5.15: A flowchart for ML estimator optimization. Here the EQRF value provides feedback for a lower bound and ML provides an upper bound. Finding the value which minimizes the squared error provides robust output.

performance. From Figure 5.18 and Figure 5.19, there also appears to be a lower k value and an upper k value that describe the GPD which fits the data.

For cluster 1, using only the most extreme data points (ends of the tail/ small u), initially indicates the upper k value. At about $u = 1,000$, the lower k value is indicated. The subset is roughly one-tenth of the total data set size, and thus the majority of the tail of the distribution is taken into account. As u increases to incorporate more of the body of the distribution, the upper k value is approached again.

This result indicates that cluster 1 may be modeled by more than one GPD. Cluster 1 is the larger of the two analyzed in this work, with 21,384 pixels. Therefore, it is reasonable that there is more than one process involved in the material content. As in [54] and [55], a mixture of GPDs, one for the body of the data and one for the tail, should be implemented. Cluster 2 is the smaller of the two with 11,557 pixels. From Figure 5.19 it can be seen that the initial tail extremity k value is large.

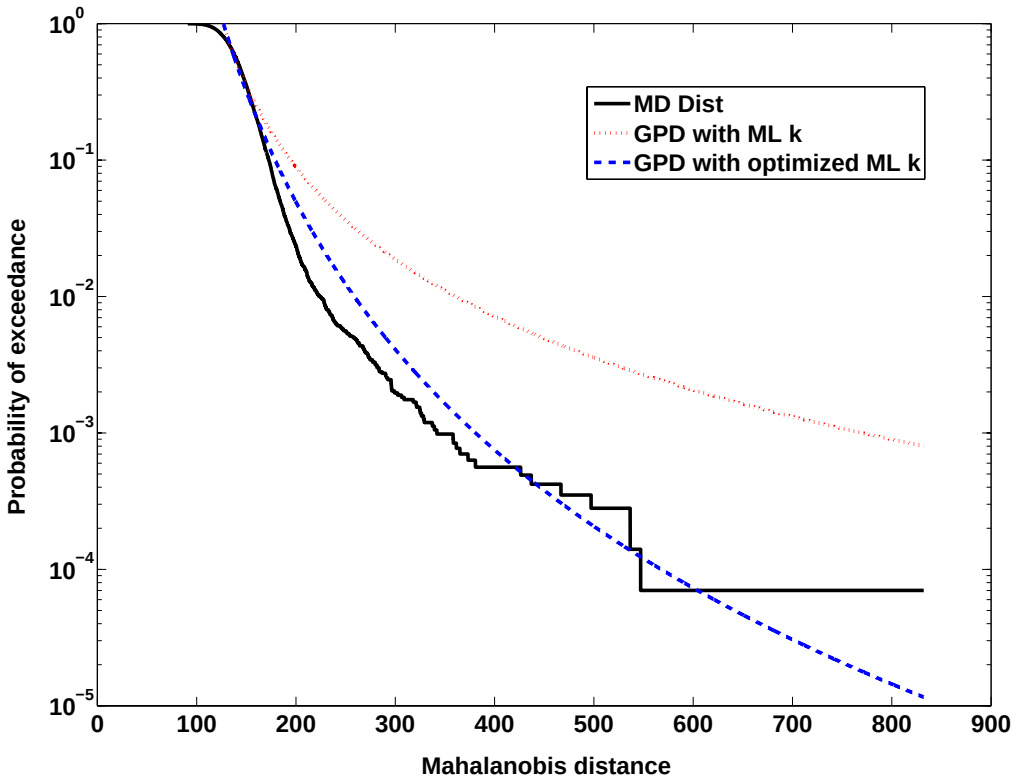


Figure 5.16: An exceedance plot of the HSI MD data (solid lines) from the cluster in Figure 5.8, the initial GPD fit to the data with the ML estimate ($u = 1,000$) for k (thin dotted line), and the second-pass optimized ML estimated k GPD fit (thick dotted line). The MSE for the initial fit is 2.1E-03 and the MSE for the optimized fit is 7.2E-06.

However, as u increases k for the entire distribution appears to converge to a single value. Here, a single GPD model of the data is sufficient. Further work on mixtures of GPDs that model HSI data clusters is needed using the framework developed here.

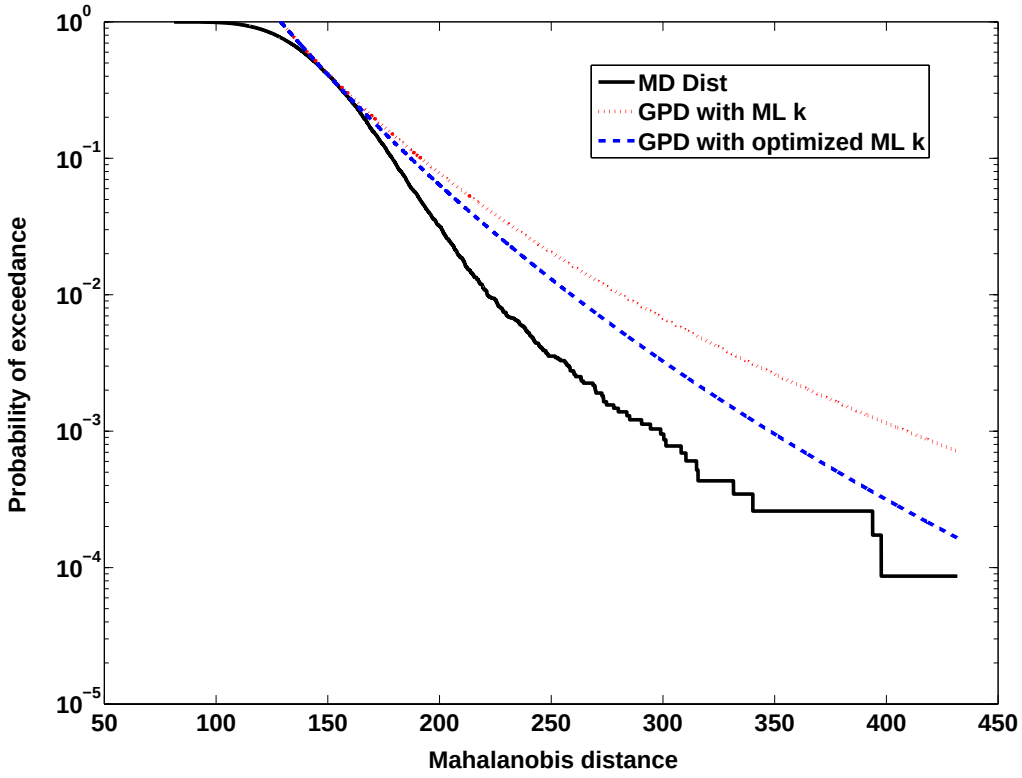


Figure 5.17: An exceedance plot of the HSI MD data (solid lines) from the cluster in Figure 5.9, the initial GPD fit to the data with the ML estimate ($u = 1,000$) for k (thin dotted line), and the second-pass optimized ML estimated k GPD fit (thick dotted line). The MSE for the initial fit is 3.0E-03 and the MSE for the optimized fit is 7.5E-04.

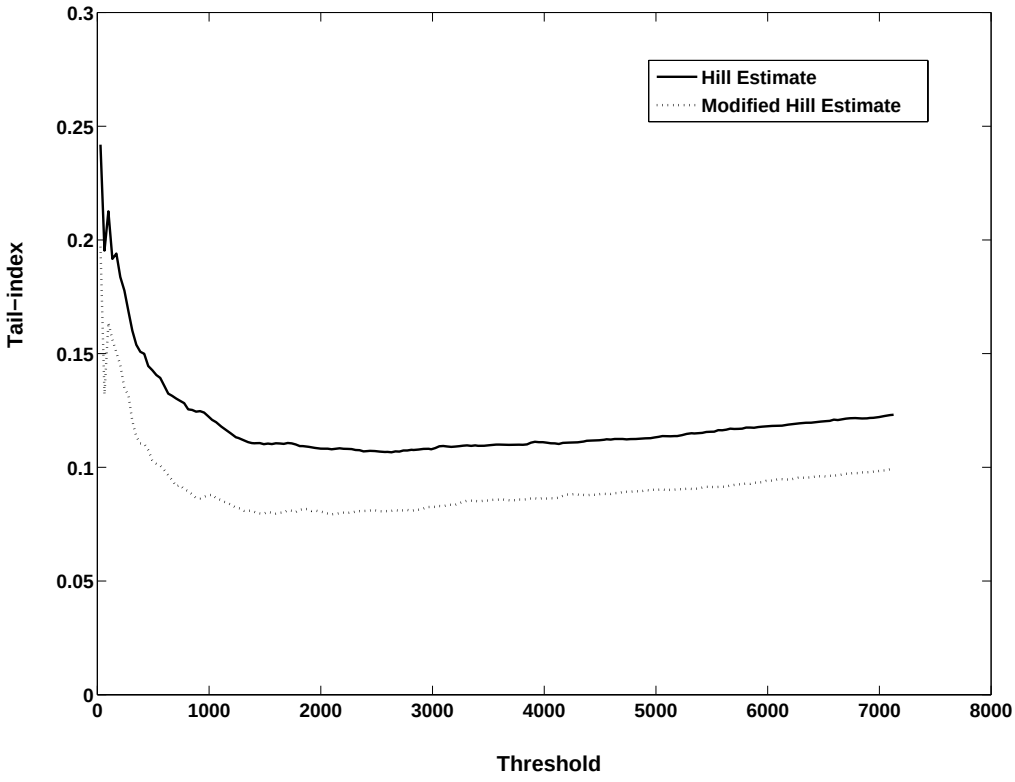


Figure 5.18: Classic Hill estimator and two-pass optimized Hill estimator performance with respect to changing the threshold u for the cluster in Figure 5.8.

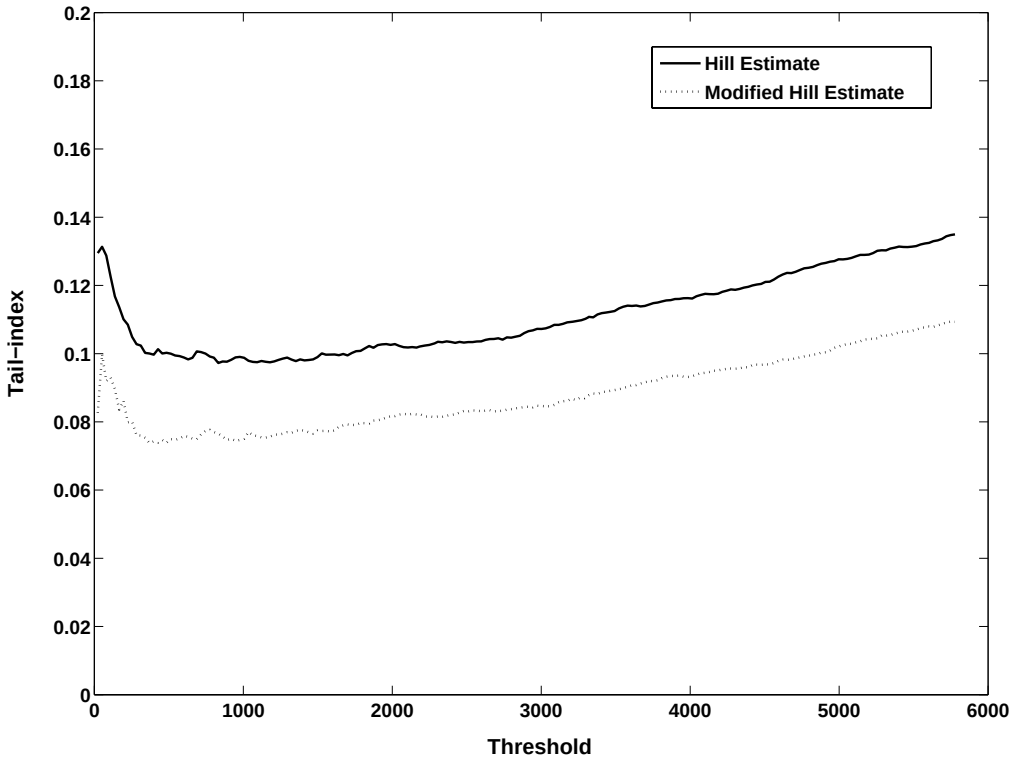


Figure 5.19: Classic Hill estimator and two-pass optimized Hill estimator performance with respect to changing the threshold u for the cluster in Figure 5.9.

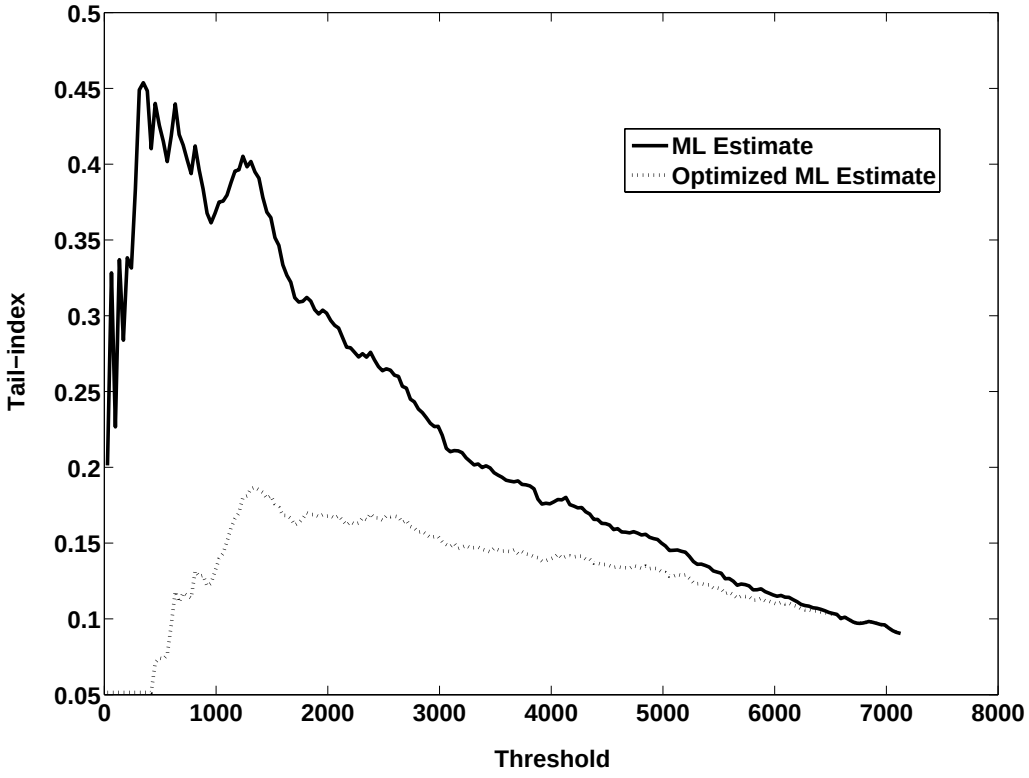


Figure 5.20: ML estimator and improved ML estimator performance with respect to changing the threshold u for the cluster in Figure 5.8.

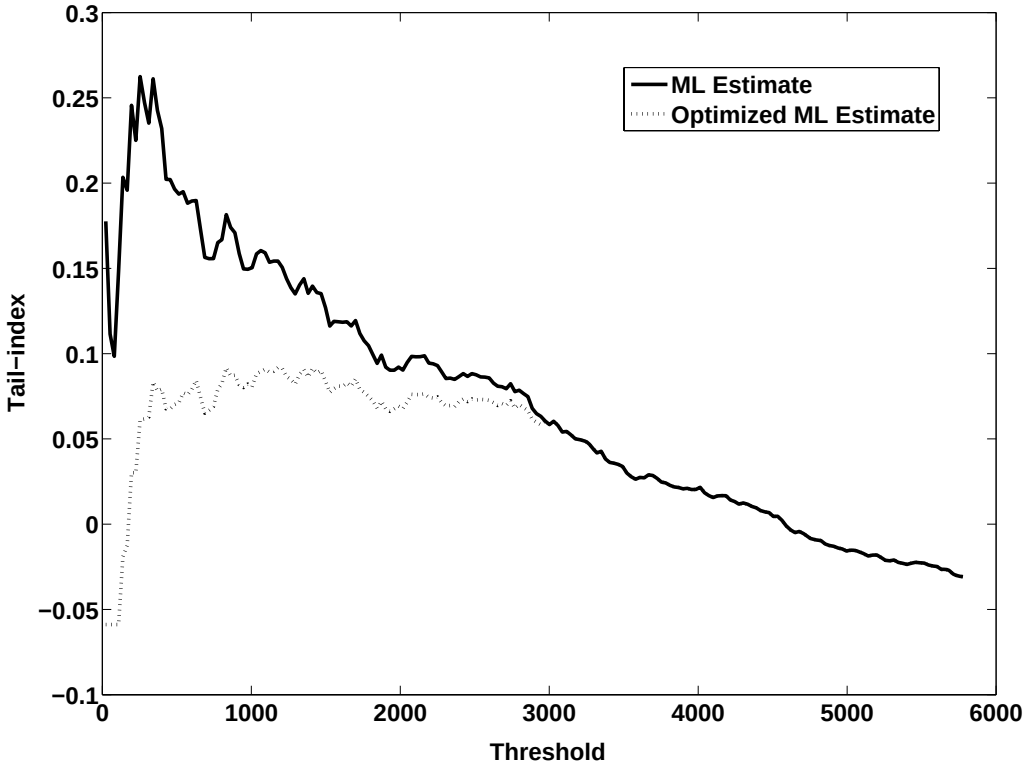


Figure 5.21: ML estimator and improved ML estimator performance with respect to changing the threshold u for the cluster in Figure 5.9.

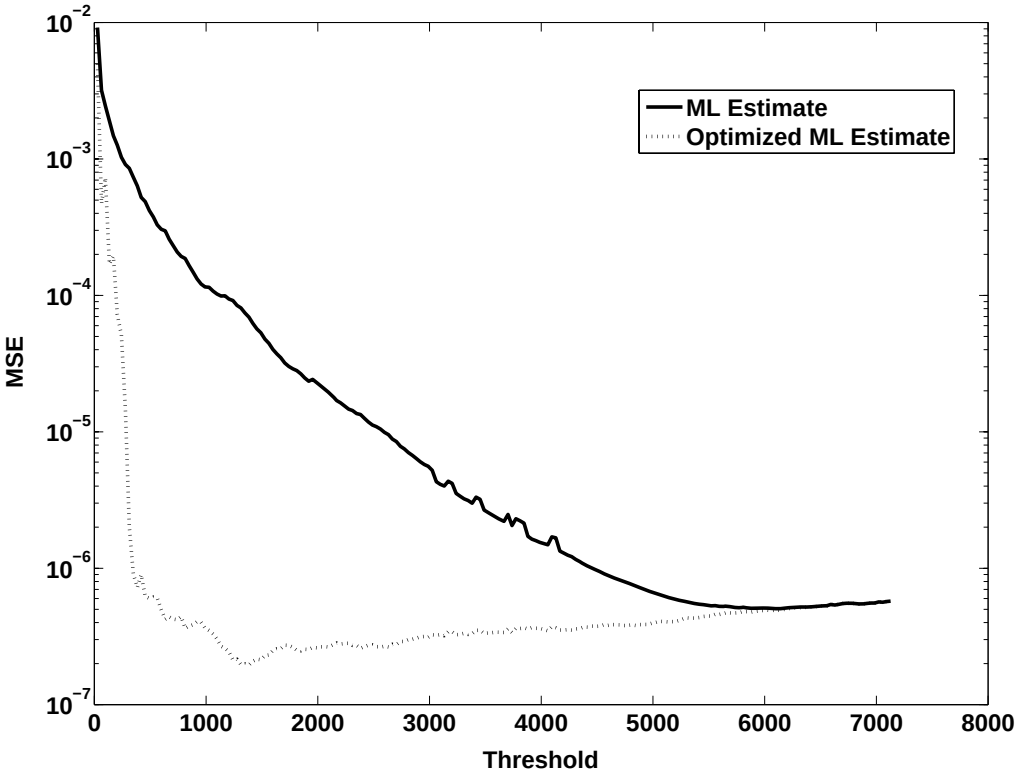


Figure 5.22: MSE of the ML estimator and improved ML estimator with respect to changing the threshold u for the cluster in Figure 5.8.

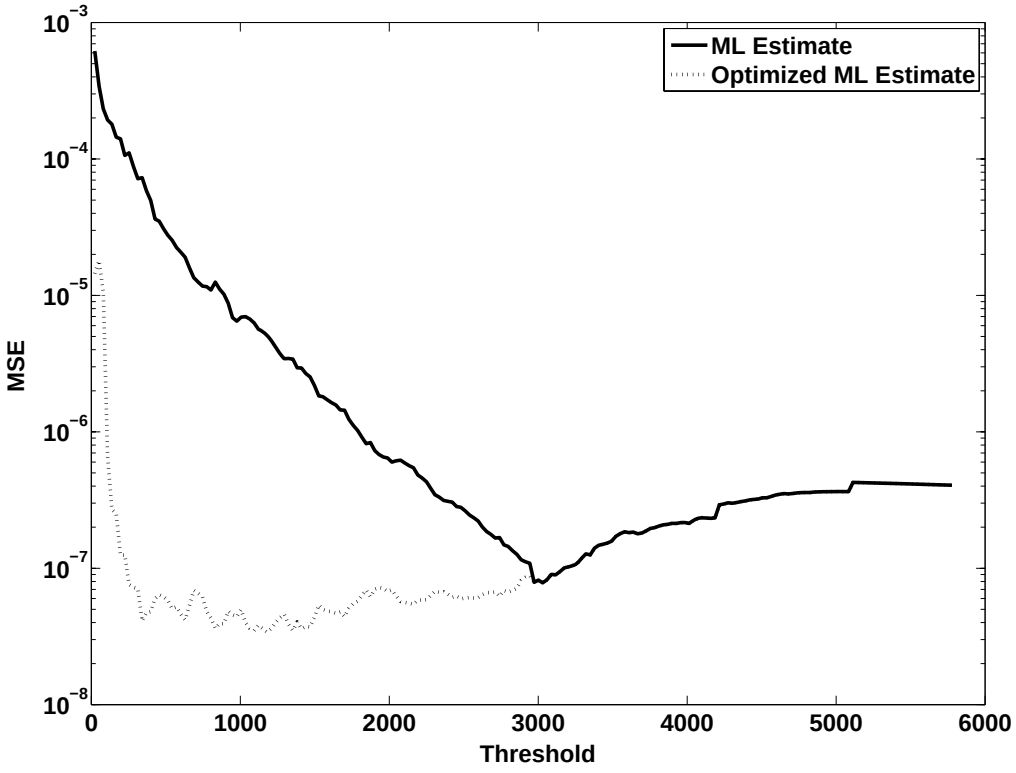


Figure 5.23: MSE of the ML estimator and improved ML estimator with respect to changing the threshold u for the cluster in Figure 5.9.

VI. Comparing Approaches and Assessing Utility

In this chapter, the MD data fitting method developed using an optimized Johnson S_L distribution is compared against the two optimized GPD fitting methods. The comparison is performed on a well-studied HSI data cube with specific regions of interest (ROIs) identified as unique background types. The different methods developed here are compared for optimality with respect to minimizing the MSE in fitting MD data, robustness to possible "outliers," and computational efficiency.

6.1 Comparison of Methods on Benchmark HSI ROIs

The HSI cube selected for performing a comparison of methods is from a Ft AP Hill image taken by the AVIRIS sensor, run 03, scene 09, on 08 November 1999. This scene is well characterized from ground information, i.e., image pixels are correlated to information in the ground spatial coverage. From the data collected on the ground, the image pixels are segmented into ROIs that exhibit similar background characteristics. For example, an ROI of an area with a particular type of vegetation is created, which mimics the process of clustering similar pixels into a specific class (as mentioned in Section 2.4). The scene is shown in Figure 6.1.

This scene is used in Chapter III. However the clusters in Chapter III are results of SEM clustering and not based on ground features. Therefore, a more involved description of the scene is not attempted in that Chapter. Also, two clusters from this scene are used in developing the optimal tail-index estimation methods for a GPD model in Chapter V. Again, explanation of the utility of ground information ROIs is not necessary in that Chapter. In this analysis, a description of the content of the ROIs is necessary to explain the variability in the scene and the performance of different methods for fitting the MD data.

In the next Section each ROI is analyzed by collecting the MD data and then fitting the distribution of the MDs using a mixture of F-distributions, the Johnson S_L distribution, and a GPD with two optimized estimators for the tail-index parameter. Each separate ROI is briefly described and an analysis performed using the four

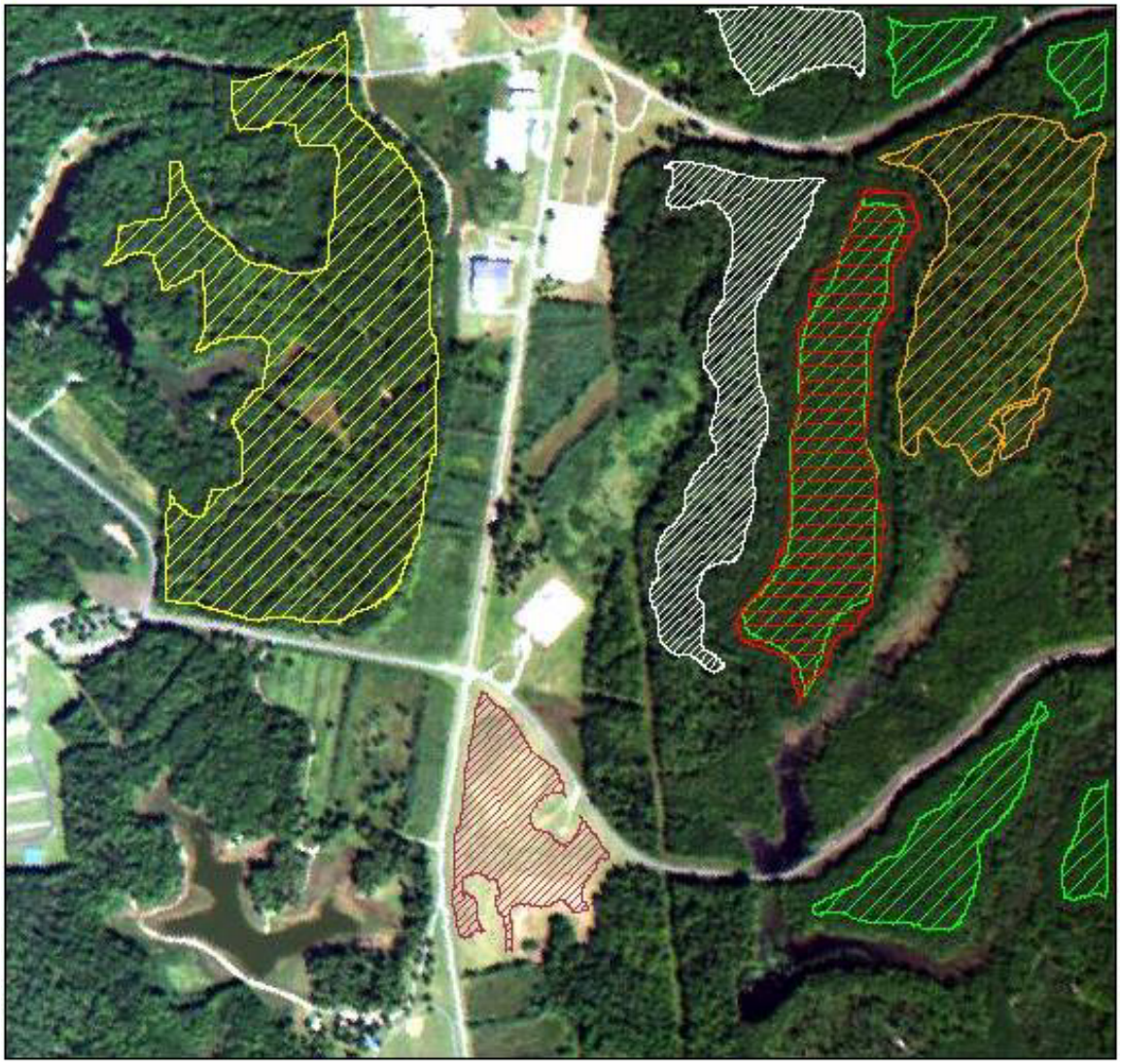


Figure 6.1: The benchmark ROIs (clusters) from the Ft AP Hill AVIRIS data collect highlighted by the masks over each area of interest. In the next Section each ROI is described and then analyzed by the methods developed here for fitting MD data.

methods. A probability of exceedance plot for each method on the ROI is then shown as an initial result. In the Section following the performance of each fit is tabulated, further results are given, and comments made about each analysis.

The first ROI is the South Panels Field Cropped (SPFC) region, shown in Figure 6.2, which contains 4,466 pixels. The variability in this region is depicted in the plot of mean pixel spectrum, standard deviations and upper/lower spectral extremes in Figure 6.3. This Figure is generated using the ROI statistics tool found in the ENVI software, which is a hyperspectral image analysis software suite [76].

Figure 6.2: The ROI of a field of grass at the southern end of the image. The ROI also contains pixels mixed with dirt, grass, and small rectangular panels. This mixture of different pixels in the ROI creates the variability noticed in Figure 6.3

The MD data are fit using the mixture of F-distributions, the Johnson S_L distribution, and a GPD with two optimized estimators for the tail-index parameter. The result is displayed in Figure 6.4. Tabulated results are given in the next Section.

The next ROI is the Mixed Forest Cropped (MFC) region, shown in Figure 6.5, which contains 9,157 pixels. The variability in this region is depicted in the plot of mean pixel spectrum, standard deviations and upper/lower spectral extremes in Figure 6.6.

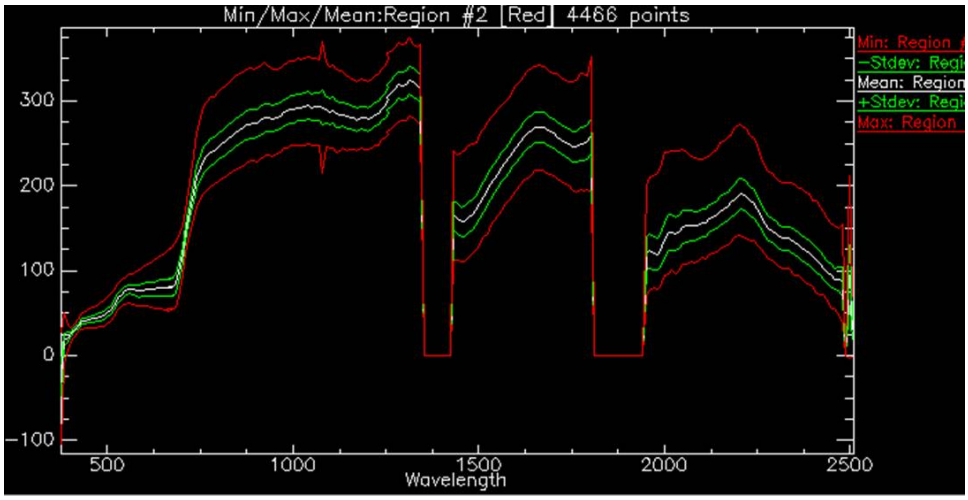


Figure 6.3: The spectral variability for the ROI in Figure 6.2. The y-axis values are units of spectral reflectance. The middle spectrum is the mean, the adjacent spectra above and below represent one standard deviation, and the top and bottom spectra are the minimum and maximum in magnitude.

The remaining six ROIs are shown along with their spectral statistics and exceedance plots resulting from the application of the methods described in Appendix F . Each ROI is briefly described in the Figures. The ROIs are each abbreviated as:

- Mixed Coniferous Forest Cropped (MCFC)
- Deciduous Forest Cropped (DFC)
- Coniferous Forest Cropped (CFC)
- All Loblolly Pine Plantations Cropped (ALPPC)
- Coniferous Forest Cropped Reduced (CFCR)
- All Loblolly Pine Plantations Cropped Reduced (ALPPCR)

The tabulated performance of each fitting method is given in the next Section.

6.3 Tables of Results and Comments

The results from the analysis described in the previous Section and the output shown in Appendix F are detailed in the following tables, where a separate table for each ROI is given that lists the performance metrics for each method. Computational

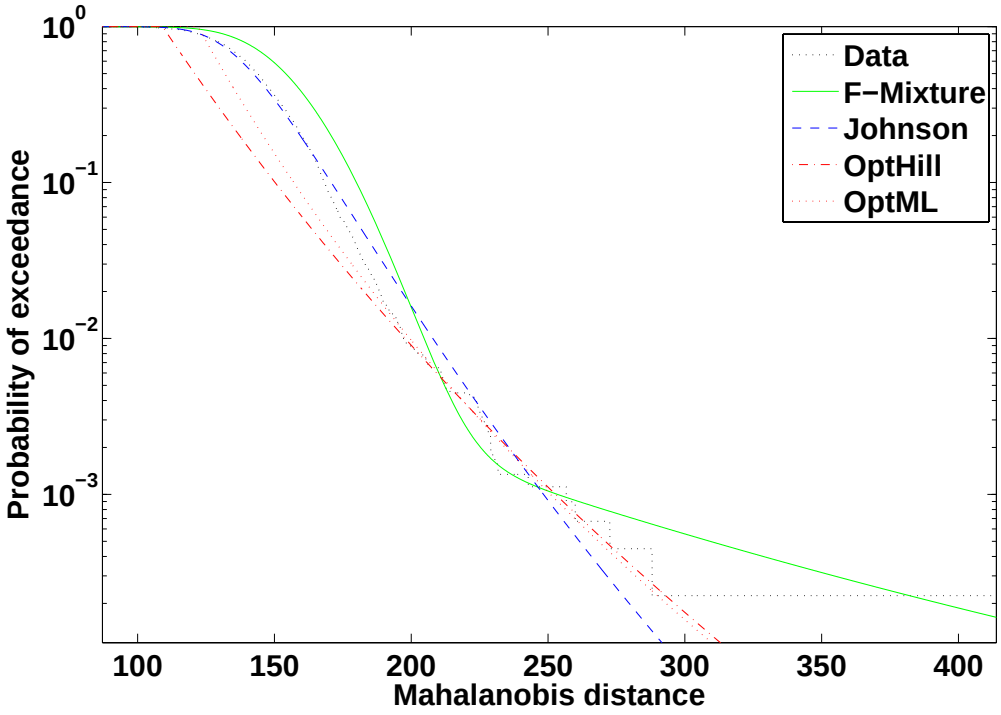


Figure 6.4: Probability of exceedance plot showing the performance of each method used to fit the MD distribution from the SPFC ROI. Notice how the F-mixture follows all the most extreme data points. The MSE and weighted MSE are given in tabulated form in the next section.

time is based on each routine programmed and executed in Matlab® version 7.1 on a conventional Pentium 4 processor. A synopsis of the findings from the tables is given at the end of this section. These findings lead to discussions and conclusions on the optimal method for HSI processing of MD data in the next section.

In comparison to Chapter III and Chapter V, the MSE and weighted MSE values are different in this Chapter due to the way the metrics are computed. In the analysis in this chapter the MSE and weighted MSE are computed in the region 10^{-2} to the smallest value on the y-axis on the probability of exceedance plot for the data (which results in fitting the tail of the data set), and are not computed over the entire data set (as in Chapter III and Chapter V). Because the interest is in fitting the tails of the data. Therefore, the MSE and weighted MSE values are different in this chapter.

Figure 6.5: The ROI of an assortment of tree types. The variability in this cluster of pixels is shown in Figure 6.6. The MD data from this cluster are fit with a mixture of F-distributions, a Johnson S_L distribution, and a GPD with two optimized estimators for the tail-index parameter. Result are displayed in Figure 6.7.

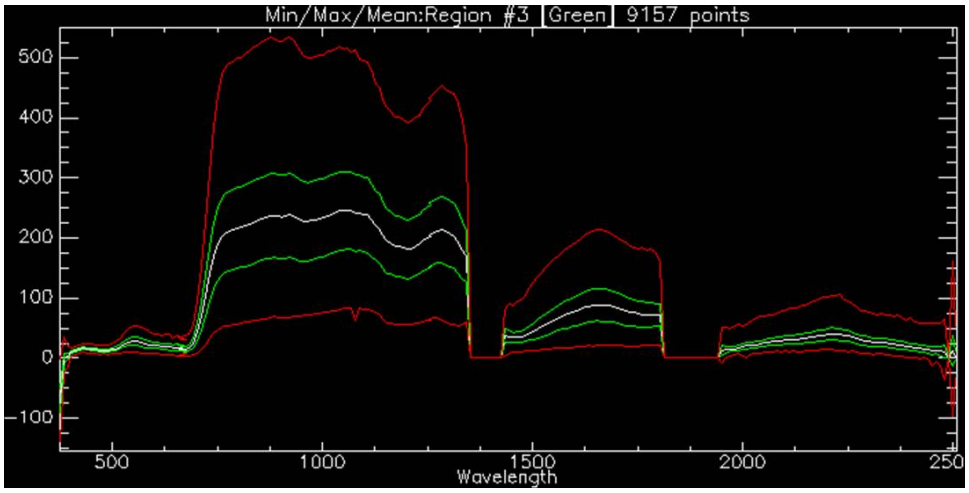


Figure 6.6: The spectral variability for the ROI in Figure 6.5. The y-axis values are units of spectral reflectance. The middle spectrum is the mean, the adjacent spectra above and below are one standard deviation, and the top and bottom spectra are the minimum and maximum in magnitude.

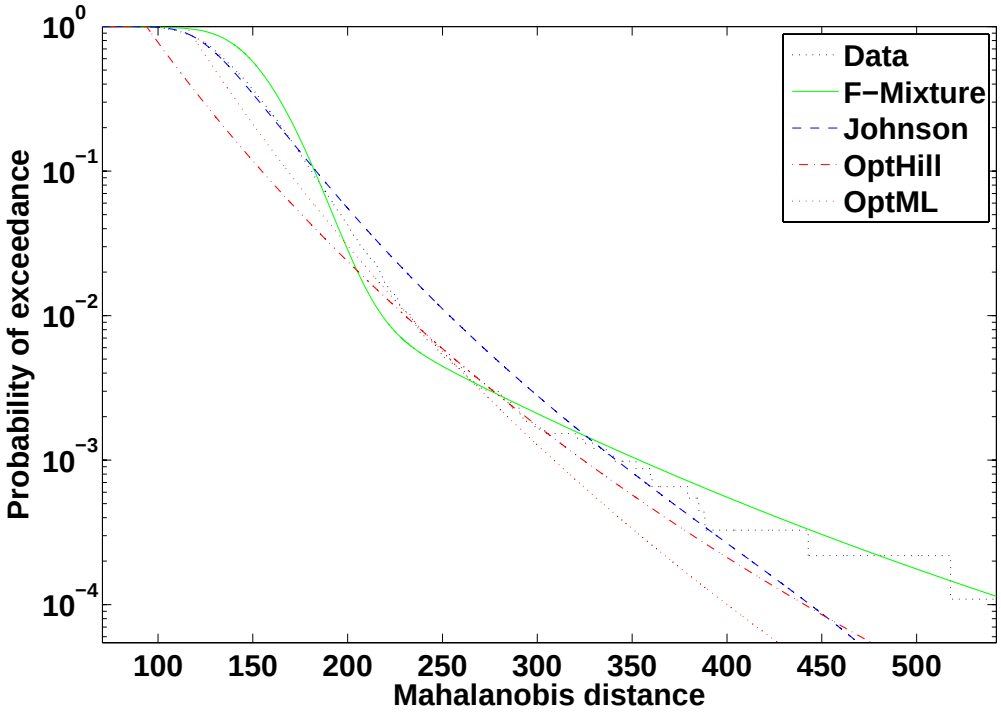


Figure 6.7: Probability of exceedance plot showing the performance of each method used to fit the MD distribution from the MFC ROI. Notice how the F-mixture follows all the most extreme data points. The MSE and weighted MSE are given in tabulated form in the next Section.

Also, in order to obtain the best possible fit for the GPD fitting methods, the cutoff $u = 1000$ used in Chapter V is not used in this comparative analysis. Instead, $u = 0.2 \cdot \text{size(MD)}$ is used for tail-index estimators for GPD models. This choice increases the performance of the tail-index estimator, in accord with the estimator threshold sensitivity analysis in Chapter IV. As a result, some of the exceedance plots of the GPD fit change with respect to the exceedance plots in Chapter V.

From the tables, it is very clear that the GPD fitting methods with the improved tail-index estimators are the most computationally efficient. The Johnson method is the fastest by far, but yields less than optimal results (MSE and weighted MSE). The F-Mixture method yields results comparable to the GPD fitting but requires much more computational time.

Table 6.1: Summary of performance for SPFC MD Data (ROI = 4,466 pixels).

Method	Computation Time (seconds)	MSE ($\times 10^{-6}$)	Weighted MSE ($\times 10^{-4}$)
F-distribution Mixture	183.05	1.78	7.63
Johnson S_L Distribution	0.06	2.28	9.83
GPD with optimized-Hill k	43.94	0.07	0.33
GPD with optimized-ML k	12.65	0.09	0.44

Table 6.2: Summary of performance for MFC MD Data (ROI = 9,157 pixels).

Method	Computation Time (seconds)	MSE ($\times 10^{-6}$)	Weighted MSE ($\times 10^{-4}$)
F-distribution Mixture	185.24	0.27	25.10
Johnson S_L Distribution	0.06	4.51	34.31
GPD with optimized-Hill k	84.11	0.06	7.03
GPD with optimized-ML k	22.50	0.14	16.04

Table 6.3: Summary of performance for MCFC MD Data (ROI = 23,411 pixels).

Method	Computation Time (seconds)	MSE ($\times 10^{-7}$)	Weighted MSE ($\times 10^{-2}$)
F-distribution Mixture	189.87	0.65	1.83
Johnson S_L Distribution	0.06	0.61	2.08
GPD with optimized-Hill k	134.64	0.25	1.30
GPD with optimized-ML k	62.94	1.69	5.30

Table 6.4: Summary of performance for DFC MD Data (ROI = 11,557 pixels).

Method	Computation Time (seconds)	MSE ($\times 10^{-6}$)	Weighted MSE ($\times 10^{-4}$)
F-distribution Mixture	187.28	0.24	1.11
Johnson S_L Distribution	0.06	1.57	7.28
GPD with optimized-Hill k	113.14	0.01	0.07
GPD with optimized-ML k	28.62	0.03	0.18

For the MSE and weighted MSE, in a majority of cases, the GPD fit with optimized tail-index estimators outperforms the other two methods by an order of

Table 6.5: Summary of performance for CFC MD Data (ROI = 9,212 pixels).

Method	Computation Time (seconds)	MSE ($\times 10^{-6}$)	Weighted MSE ($\times 10^{-4}$)
F-distribution Mixture	185.81	0.07	0.25
Johnson S_L Distribution	0.06	3.19	9.96
GPD with optimized-Hill k	79.97	0.05	0.17
GPD with optimized-ML k	22.13	0.06	0.19

Table 6.6: Summary of performance for ALPPC MD Data (ROI = 14,257 pixels).

Method	Computation Time (seconds)	MSE ($\times 10^{-6}$)	Weighted MSE ($\times 10^{-4}$)
F-distribution Mixture	187.76	3.27	23.03
Johnson S_L Distribution	0.07	0.49	20.99
GPD with optimized-Hill k	129.22	0.36	19.10
GPD with optimized-ML k	38.16	0.21	14.60

Table 6.7: Summary of performance for CFCR MD Data (ROI = 8,533 pixels).

Method	Computation Time (seconds)	MSE ($\times 10^{-6}$)	Weighted MSE ($\times 10^{-4}$)
F-distribution Mixture	183.15	0.42	11.80
Johnson S_L Distribution	0.05	0.08	2.43
GPD with optimized-Hill k	118.92	0.02	0.85
GPD with optimized-ML k	23.96	0.03	1.01

Table 6.8: Summary of performance for ALPPCR MD Data (ROI = 12,976 pixels).

Method	Computation Time (seconds)	MSE ($\times 10^{-6}$)	Weighted MSE ($\times 10^{-4}$)
F-distribution Mixture	188.43	0.07	0.18
Johnson S_L Distribution	0.06	2.25	4.60
GPD with optimized-Hill k	121.01	0.01	0.22
GPD with optimized-ML k	32.72	0.11	1.13

magnitude or greater. In the cases where the MSE metric shows close scores, the weighted MSE provides a good discriminator, offering a scale which takes into account the greater importance of fit in further regions of the tail. For example, in the DFC ROI summary in Table 6.4, the GPD with both optimized tail-index estimators achieves low MSE values for each. However, the weighted MSE shows that, in this case, the optimized ML estimation method results in a goodness-of-fit metric almost three times better. Since the interest here is in fitting the entire tail of the distribution (not the body of the distribution or only the most extreme portions of the tail of the distribution), the weighted MSE is valuable because it provides extra information regarding deviations from the trend of the tail.

For the "trend" of the tail, the F-mixture and optimized GPD fitting methods provide good models. The Johnson S_L method does not follow the trend of the tail (it tends to model the majority of the available data points and then abruptly falls off at the furthest region of the tail), which leads to a poor MSE score and an even worse weighted MSE. However, as opposed to the Johnson S_L behavior, the F-mixture follows the trend of the tail too closely, causing it to overcompensate for the furthest extremes of the tail at the expense of other regions in the tail.

This overcompensation by the F-mixture is detrimental to robust performance. As noticed in Figure 6.7, the F-mixture takes extreme swings to fit the furthest points in the tail, which results in a poor fit to the rest of the tail. In comparison, the GPD methods with optimized tail-index estimators follow the trend of the tail but do not shift upward to match the most extreme data points. Given a ROI where background pixels are mixed with anomalous target pixels, the F-mixture attempts to fit the anomalous pixels, while the optimized GPD follows the trend of the background data and clearly indicates anomalous pixels as those lying above and to the right. This type of behavior is also seen in the ALPPC ROI exceedance plots and the table data.

6.4 Summary of Results

This analysis identifies the most desirable method for fitting a distribution of MD data from an HSI cube. The different methods are compared for optimality with respect to minimizing the MSE in fitting MD data, robustness to possible "outliers," and computational efficiency. It is clear from the figures, tables, and discussion that *GPD methods with optimized tail-index estimators are preferred over the Johnson S_L distribution and F-mixture fitting techniques.*

These improved GPD fitting methods are developed for the generic case of fitting a distribution which exhibits possible outliers (possible targets in a target plus background scenario) in the data set. They may be further developed for robustness and optimality in different scenarios. For example, if a mixture of two GPDs fit to a MD distribution is desired, one tail-index estimation technique for a GPD may be modified to include possible outliers (one developed to fit all of the points in the extremities of the tail) and another modified to exclude a greater number of possible outliers (one tailored to fit the body). The mechanics for such modifications are identified and described in the previous Chapter. For this generic case (a single distribution model fits the data) these improved tail-index estimators for GPDs perform best given the specified criteria and results.

In conclusion, this chapter presents the final phase of research and discusses the last of five research objectives. The capstone is the identification and validation of an optimal method for modeling MD data distributions from an HSI cube. A review of all research objectives is given in the next chapter along with a final summary, review of the contributions, and recommendations for future work.

VII. Summary and Conclusions

This research has developed methods that improve models of MD distributions from HSI data. Accurate models of MD behavior are important for achieving reliable metrics in post-processing routines and setting accurate thresholds for robust detector performance. A summary of the work conducted here is presented below through a review of objectives, a highlight of contributions, and a discussion of recommendations for future research.

7.1 Summary of Results

Objective 1: *Determine an optimal Johnson distribution model for HSI MD data.* The Johnson S_L distribution was determined to model MD distributions well. Because it estimates its parameters directly from quantile shifts in the data, it is susceptible to perturbations, where perturbations are recognized as possible "outliers" and/or anomalous data that cause a shift in the rate of change in the slope of the exceedance plot of the data. This model was optimized against the perturbations by mitigating the effect of these "bumps" in the exceedance plot, thus making the model more robust to data corruption.

Objective 2: *Determine a multivariate elliptically contoured distribution model from the univariate Johnson distribution.* A multivariate EC model was derived from the univariate Johnson distribution. This model was developed using EC theory, and the derivation was compared to that of the multivariate t-distributed EC model from the univariate F-distribution of MD. Also, the final form of the EC density function was found to be similar to other well-known multivariate forms which result in heavy-tailed MD distributions.

Objective 3: *Determine a viable parameter estimation method for obtaining tail-index parameters for GPD models.* Different parameter estimation methods were analyzed for fitting a GPD to MD data. These methods were tested with respect to suitability for HSI MD data models and to sensitivity of the estimators to data quantity. The most robust estimation method was determined to be the ML estimator,

along with an approximation to the ML estimator, the Hill estimator. The most effective data subset threshold was found to be 20 % of the HSI cluster size.

Also, a posterior density estimation method was developed using Bayesian techniques with a gamma distributed prior for k. This method incorporates Parzen window density synthesis to facilitate estimation from only a few points. The result is a method from which posterior metrics can be derived in a computationally efficient manner from a few guesses at k values for a GPD model of a data set.

Objective 4: *Develop an optimal method for obtaining robust tail-index estimates for GPD models.* The ML estimation and Hill estimation methods were optimized for minimal MSE with respect to possible outliers and perturbations in data. An automated, robust, and computationally efficient algorithm was developed for the GPD parameter estimation method that is suitable for HSI data processing. The algorithm provides options to the optimization process so that it can be modified to change the performance of the estimator. As a result of simulations performed on similar data, these options were set to values that yield optimal results given HSI MD data distributions from vegetative clusters. With similar experimentation, different cluster types can easily be accommodated.

Objective 5: *Assess the utility of Johnson and GPD models for stochastic HSI data processing.* The Johnson model was compared against the GPD model, and the estimation methods and their robustness under different data configurations are assessed. A comparison was performed on a well-studied HSI data cube with specific ROIs identified as unique background types. The different methods developed in this research were compared for optimality with respect to minimizing the MSE in fitting MD data, robustness to possible "outliers," and computational efficiency. The results of this comparison are shown in Figure 7.1.

Based on these results, the optimized tail-index estimation routines for GPD fitting to MD data were found to be best for efficiently achieving greater accuracy in stochastic HSI data processing.

	Optimized GPD	Johnson S_L	*F*-mixture
Computaional Efficiency	Optimal Hill estimator generally requires ~ 40 – 130 seconds	Requires less than one second	Requires more than 180 seconds
Minimal MSE	Results in minimal MSE in tail region	Results in less than minimal MSE in tail region	Results in minimal MSE in tail region
Robustness to "outliers"	Optimized for robustness against possible "outliers"	Optimized for robustness against possible "outliers"	Models all data points, including possible "outliers"

Figure 7.1: Matrix of results from the comparative analysis. The diagonally hatched boxes represent undesirable performance, the dotted box represents less than optimal performance, and the white boxes represent optimal performance. Notice that no column has all white entries. However, the optimized GPD model results are most favorable compared to the other columns.

7.2 Contributions

This research advances HSI data processing and exploitation. The following are unique contributions of this research:

- *Determined a means of mitigating possible "outlier" effects to Johnson S_L distribution models of MD data.* The application of Johnson S_L distributions to a data set is not new, however, using the second derivative of the exceedance plots of the S_L model to mitigate the effects of possible "outliers" on the model fit is unique to the field.
- *Determined a form for the multivariate data density function that generates MD data distributed according to a Johnson S_L distribution.* The EC model

developed to describe the multivariate density function that generates the MD distributions with given Johnson S_L distribution parameters is unique and is not in the literature associated with Johnson distributions and/or multivariate EC distributions.

- *Identified a tail-index parameter estimation technique that is optimal for HSI MD data processing.* Although modeling MD data with GPDs has been considered in the literature, research to determine the best type of tail-index estimator has not been accomplished prior to this work. In particular, threshold sensitivity analysis with second derivatives and posterior density estimation using Bayesian and Parzen window techniques is unique.

- *Developed a two-pass algorithm that optimizes the ML and Hill estimators of tail-index values for GPD models.* The Hill and ML estimators are optimized against the effects of possible "outliers" in order to improve the MSE of the GPD fit to MD distributions. This optimized technique is a new and unique algorithm for robust HSI MD processing, and it is implemented to demonstrate improved performance.

- *Developed a tail-index parameter posterior density estimator which uses limited available data.* A Bayesian estimation process is applied which uses a uniform prior density to estimate the posterior density of the tail-index parameter from a GPD model that fits heavy-tailed data. The posterior density is then estimated using gamma Parzen window kernels. The procedure is important because it provides a capability for extracting metrics on the estimated tail-index value. For ML estimation, confidence intervals and further metrics may be derived from asymptotic properties. However, for real-world and real-time data, sample sizes rarely approach the asymptotic limit, and a method that is effective with the data available is necessary. The method described here demonstrates the development of a density that estimates the tail-index parameter of a data set (similar to HSI MD data) using only nine inputs.

7.3 Recommendations for Future Research

The multivariate EC model that generates the Johnson S_L distributed MDs was given. Further research into simulating HSI-type multivariate data, which results in MDs distributed according to the Johnson S_L distribution with the same parameters, could be pursued. This further work may be especially promising given the similarity between the form of the derived multivariate EC function and known multivariate density functions.

In analyzing the initial performance of the optimized Hill and ML estimators in Chapter V, it was noticed that there are two GPD forms that model MD distributions from HSI data. The optimized methods developed here, and methods that employ linear mixing of two F-distributions used to model MDs, could be investigated to develop more insight into using two GPDs to model the MD data. All of the parameters of the GPDs may be obtained efficiently, and the only other variable is the weighting coefficient, for which a simple search for weights between zero and one could be implemented.

Finally, because the GEVs introduced here model univariate data, an investigation into data from single HSI bands can be performed. Information on the extent of heavy (or light) tails from data in individual bands could lead to information on band selection and band rejection algorithms for HSI data processing. Also, multivariate GEV models may be developed by observing this phenomenon and the correlation between band parameters.

Appendix A. SEM Mechanics

This appendix briefly describes the mechanics of the Stochastic Expectation Maximization (SEM) algorithm used in obtaining clusters of similar material types in an HSI image. SEM is a variant of the EM algorithm where, instead of using the statistics of all available pixels in an image during the probability assignment process, statistical information from each respective cluster is used in assigning a probability of a pixel belonging to that cluster [18, 54].

Much of the mechanics of SEM are similar to EM, other than that SEM uses individual cluster statistics and EM uses all available data. However, a Gaussian model is assumed for the composition of each cluster and, therefore, the expectation step of the algorithm needs further explanation. Specifically, the mechanics describing how a pixel is determined to belong to a specific cluster, given the cluster statistics, is explained below.

Expectation Step Mechanics. The posterior probability (probability of pixel x_n given cluster k (also known as the "abundance value")) is calculated by iterating on

$$P(x_n \mid \Psi_k) = \frac{p(\Psi_k \mid x_n)P(x_n)}{p(x_n)}, \tag{A.1}$$

where x_n is the n^{th} pixel and Ψ_k represent the parameters associated with cluster k. A lowercase p represents likelihood and an uppercase P represents a probability. This expression is Bayes' rule, which may be expressed in words as

$$posterior = \frac{likelihood \times prior}{evidence}.$$

Here, *likelihood* describes the likelihood of cluster k given pixel x_n, *prior* is taken out of the cluster information from the previous iteration (i.e., number of pixels in cluster k divided by number of pixels in image), and *evidence* is the overall likelihood of the

data set. In using the finite mixture model with Gaussians describing the distribution of pixels within their respective clusters, the posterior probability is

$$P(x_n \mid \Psi_k) = \frac{\hat{\pi}_k f_d(x_n \mid \hat{\mu}_k, \hat{\Gamma}_k)}{L(x_n \mid \Psi_k)},$$

where $\hat{\pi}_k$ is the current estimate of the prior, $f_d(x_n \mid \hat{\mu}_k, \hat{\Gamma}_k)$ is the likelihood (current multivariate Gaussian model of the cluster given the current estimates of the parameters), and $L(x_n \mid \Psi_k)$ is the evidence (overall data set likelihood). Expanding this expression yields

$$P(x_n \mid \Psi_k) = \frac{\hat{\pi}_k \left|\hat{\Gamma}_k\right|^{-\frac{1}{2}} \exp\left[-\frac{1}{2}(x-\hat{\mu}_k)^T \hat{\Gamma}_k^{-1}(x-\hat{\mu}_k)\right]}{\sum_{k=1}^{M} \left|\hat{\Gamma}_k\right|^{-\frac{1}{2}} \exp\left[-\frac{1}{2}(x-\hat{\mu}_k)^T \hat{\Gamma}_k^{-1}(x-\hat{\mu}_k)\right]}. \tag{A.2}$$

Taking the logarithm of both sides,

$$\ln[P(x_n \mid \Psi_k)] = [C_k - \frac{1}{2} M D_k] - \ln(D).$$

Here C_k stands for the value in the expression $\ln\left(\hat{\pi}_k \left|\hat{\Gamma}_k\right|^{-\frac{1}{2}}\right)$, D is the expression in the denominator of Equation (A.2), and MD_k represents the MD from the test pixel to the mean of the k^{th} cluster. If there are only three clusters ($M = 3$), the denominator term is

$$\begin{aligned}
\ln(D) = \quad & \ln \quad [\hat{\pi}_1 \left|\hat{\Gamma}_1\right|^{-\frac{1}{2}} \exp\left[-\frac{1}{2}(x-\hat{\mu}_1)^T \hat{\Gamma}_1^{-1}(x-\hat{\mu}_1)\right]] + \ldots \\
\ldots + \quad & \ln \quad [\hat{\pi}_2 \left|\hat{\Gamma}_2\right|^{-\frac{1}{2}} \exp\left[-\frac{1}{2}(x-\hat{\mu}_2)^T \hat{\Gamma}_2^{-1}(x-\hat{\mu}_2)\right]] + \ldots \\
\ldots + \quad & \ln \quad [\hat{\pi}_3 \left|\hat{\Gamma}_3\right|^{-\frac{1}{2}} \exp\left[-\frac{1}{2}(x-\hat{\mu}_3)^T \hat{\Gamma}_3^{-1}(x-\hat{\mu}_3)\right]],
\end{aligned}$$

which results in

$$\ln(D) = [C_1 - \frac{1}{2}MD_1] + [C_2 - \frac{1}{2}MD_2] + [C_3 - \frac{1}{2}MD_3].$$

Also, if it is desired to find the posterior probability of the n_{th} pixel given cluster 1, then

$$\ln[P(x_n \mid \Psi_k)] = \frac{1}{2}MD_2 + \frac{1}{2}MD_3 - C_2 - C_3,$$

which yields

$$\begin{aligned} Posterior &= \exp\left(-C_{\bar{k}} + \frac{1}{2}(MD_2 + MD_3)\right) && \text{(A.3)} \\ &= \frac{\exp\left(\frac{1}{2}(MD_2 + MD_3)\right)}{\exp\left(C_{\bar{k}}\right)}, && \text{(A.4)} \end{aligned}$$

where $\bar{k}$ includes the clusters information $\left(\hat{\pi}_k \left|\hat{\Gamma}_k\right|^{-\frac{1}{2}}\right)$ not associated with the k^{th} cluster. The denominator in Equation (A.3) involves the covariance matrices of clusters 2 and 3 such that

$$\begin{aligned} \exp\left(C_{\bar{k}}\right) &= \exp\left(C_2 + C_3\right) \\ &= \exp\left(\hat{\pi}_2 \left|\hat{\Gamma}_2\right|^{-\frac{1}{2}} + \hat{\pi}_3 \left|\hat{\Gamma}_3\right|^{-\frac{1}{2}}\right). \end{aligned}$$

The behavior of this expression over a range of $\left|\hat{\Gamma}_k\right|^{-\frac{1}{2}}$ values is given in Figure A.1.

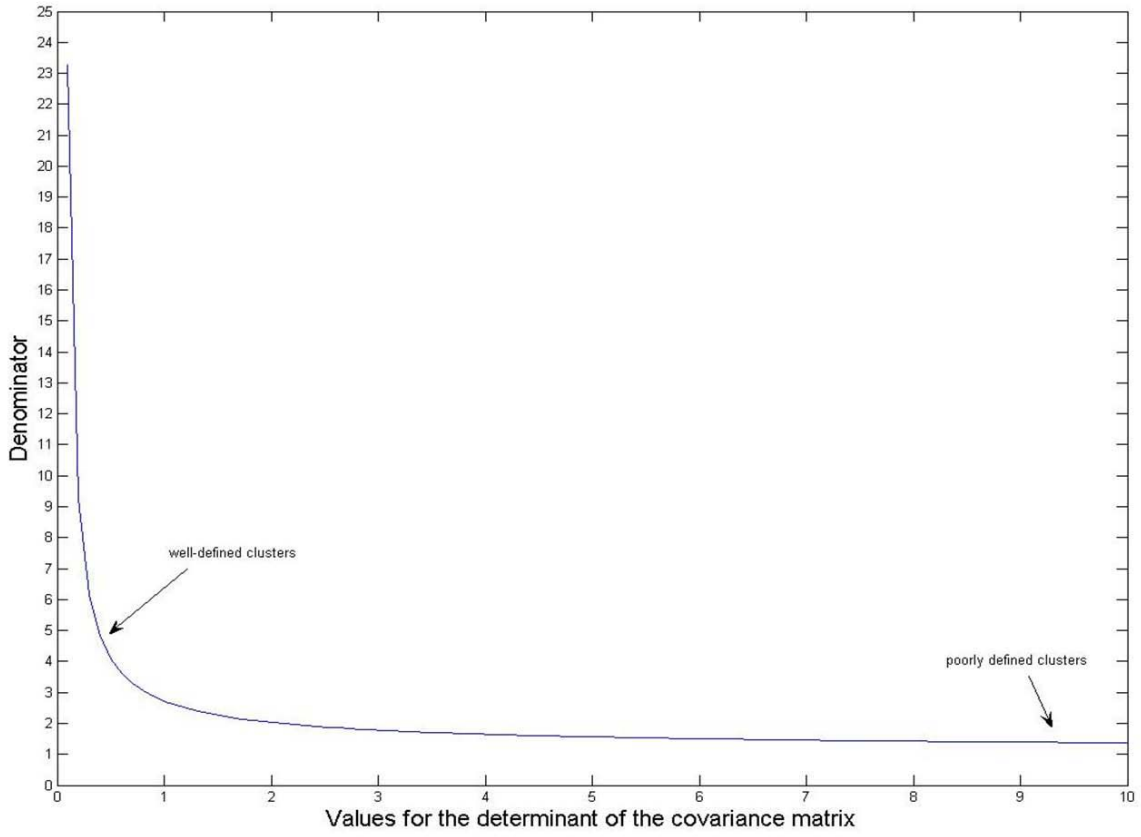


Figure A.1: Behavior of the denominator as a result of covariance matrix determinant size.

Initially, when clusters are poorly defined, the $|\Gamma_k|^{-\frac{1}{2}}$ values are larger. Therefore, the denominator for the posterior probability is smaller:

$$Posterior = \frac{\exp\left(MD_2 + MD_3\right)}{smaller\ value}.$$

Two cases are, first, if x_n belongs to cluster 1, then MD_2 and MD_3 tend to be larger, and the posterior probability expression is

$$Posterior \approx \frac{larger\ value}{smaller\ value},$$

yielding a higher value for the posterior probability, as expected. Second, if x_n does not belong to cluster 1, then MD_2 or MD_3 are smaller, and the posterior probability expression is

$$Posterior \approx \frac{smaller\ value}{smaller\ value},$$

yielding a lower value for the posterior probability, again as expected.

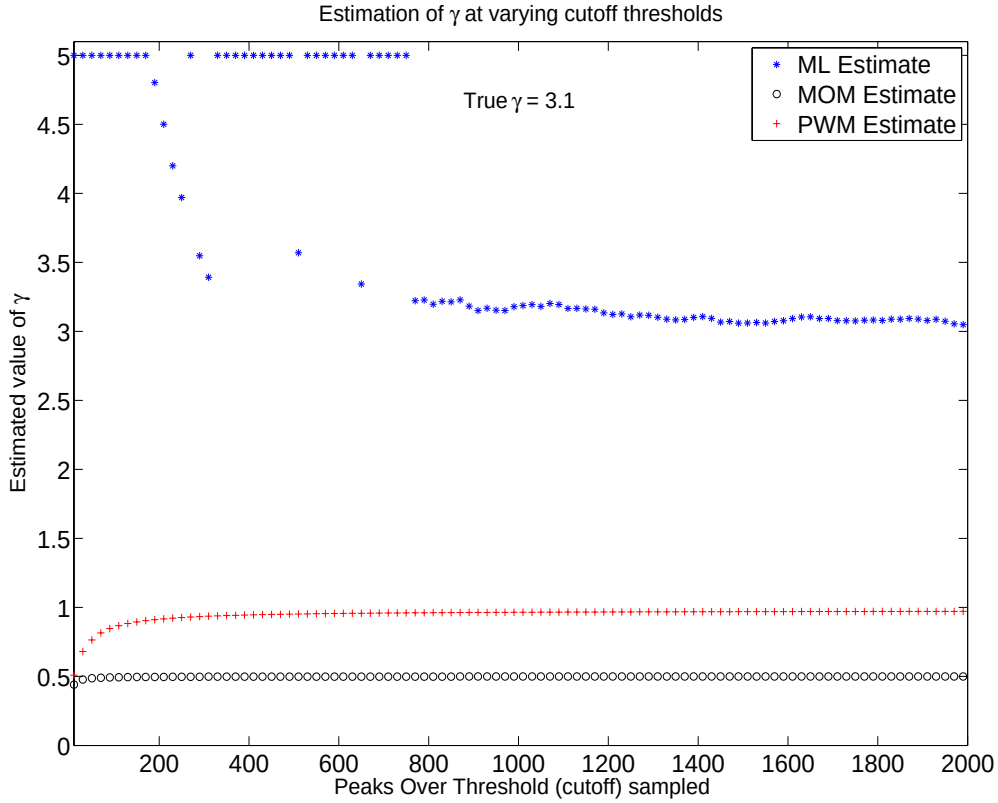
As the routine converges, clusters become more well-defined and $|\Gamma_k|^{-\frac{1}{2}}$ values become smaller. The denominator for the posterior probability is then larger, i.e.,

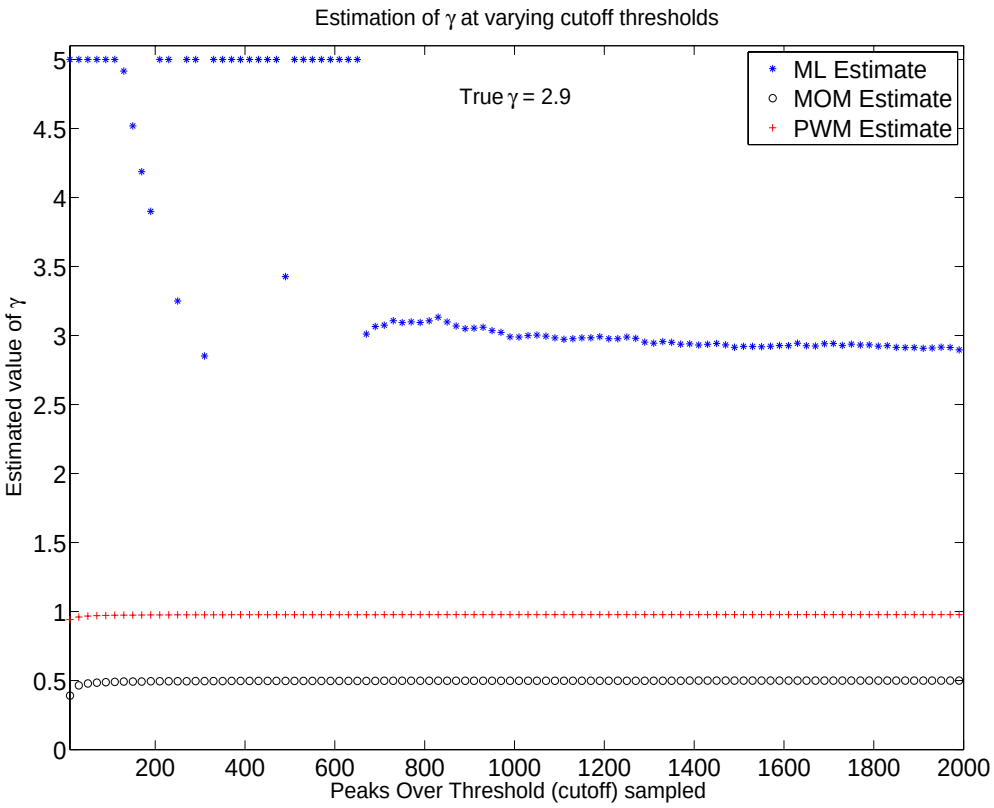
$$Posterior = \frac{\exp\left(MD_2 + MD_3\right)}{larger\ value},$$

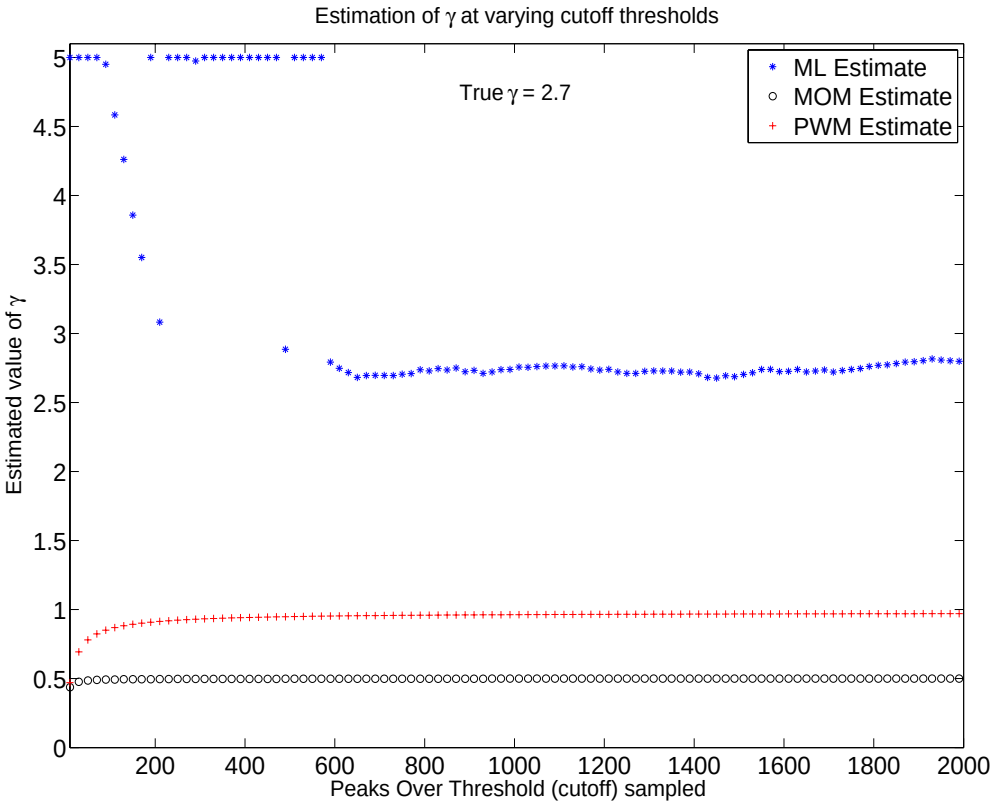
resulting in faster convergence.

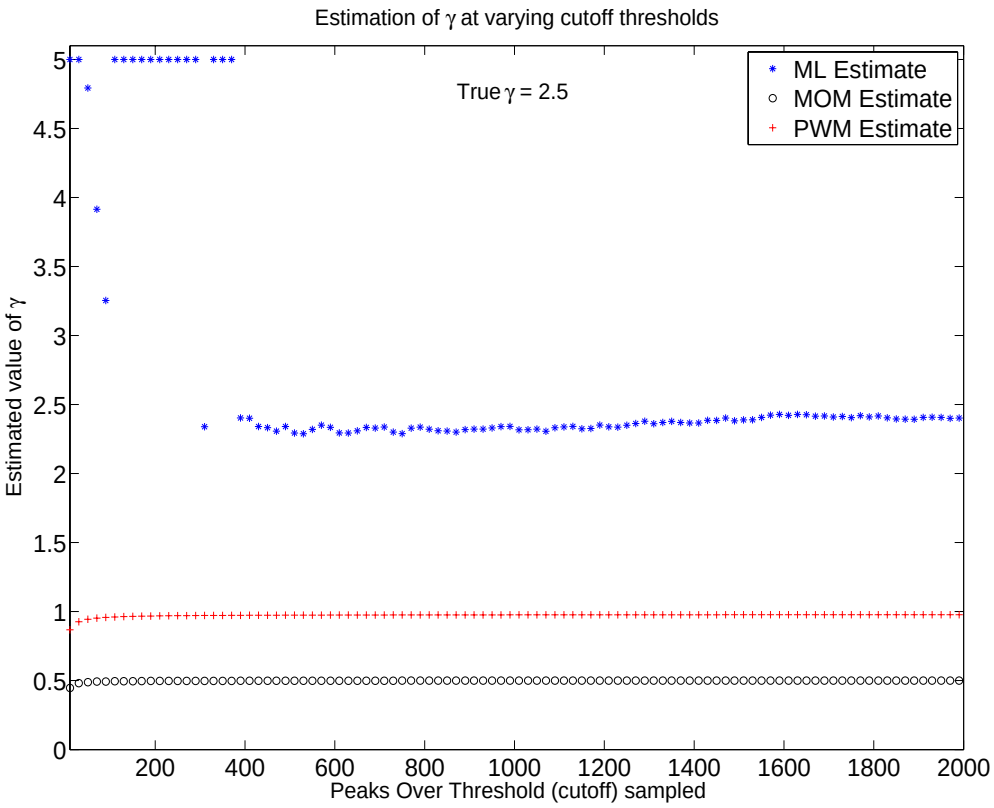
Appendix B. Simulations for ML, MOM and PWM Estimators

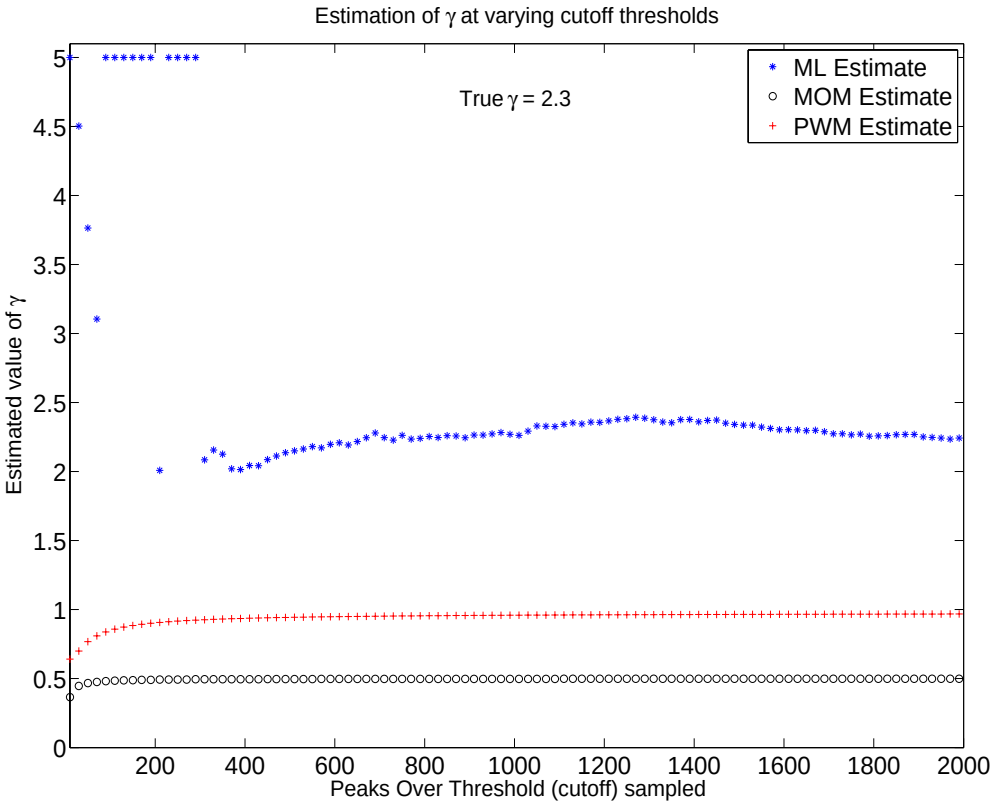
This appendix shows the results of ML, MOM and PWM estimators on simulated GPD RVs with tail-index equal to γ.

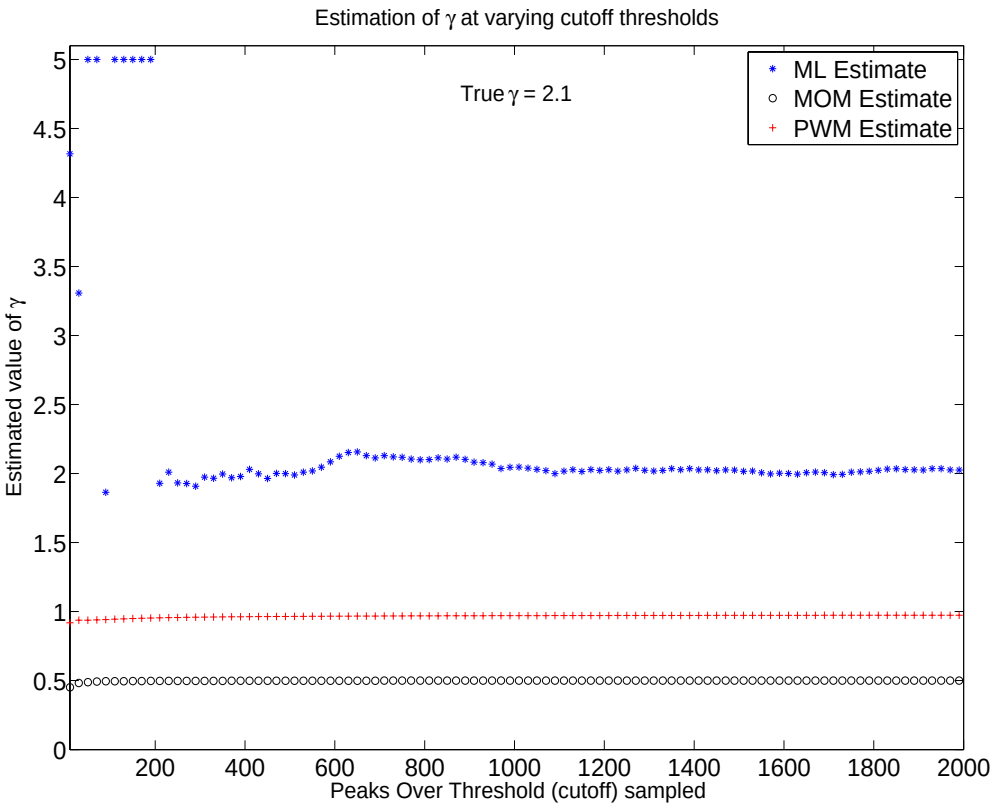


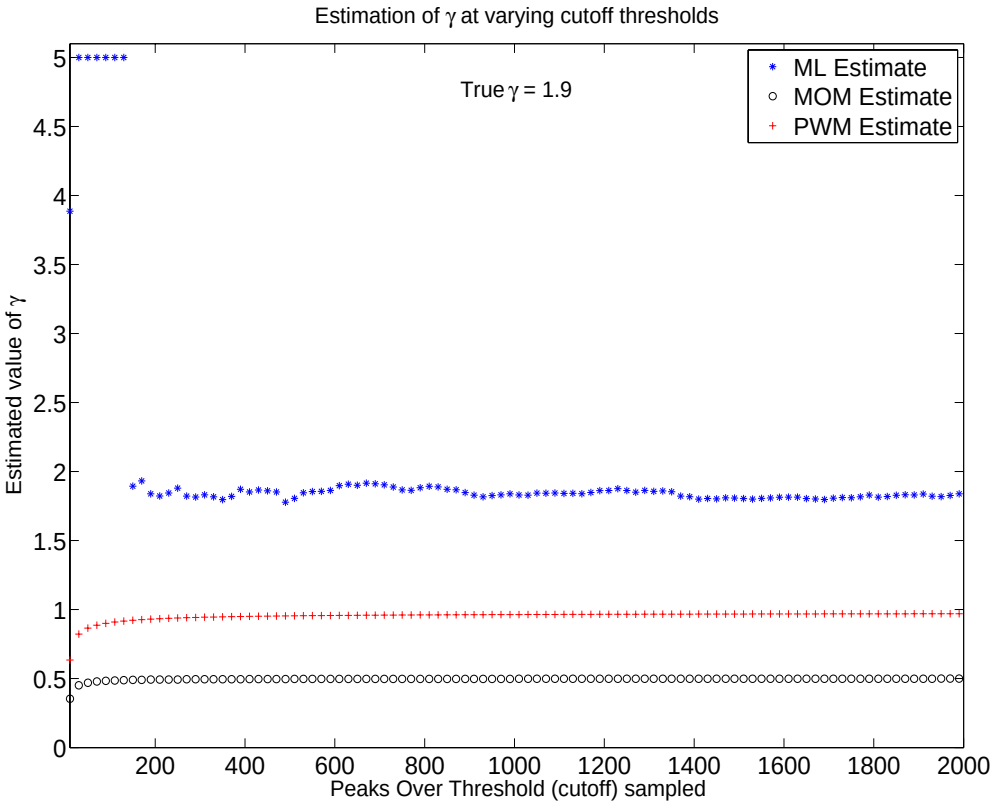

Estimation of γ at varying cutoff thresholds
True γ = 2.9
ML Estimate
MOM Estimate
PWM Estimate
Estimated value of γ
Peaks Over Threshold (cutoff) sampled
0
0.5
1
1.5
2
2.5
3
3.5
4
4.5
5
200
400
600
800
1000
1200
1400
1600
1800
2000

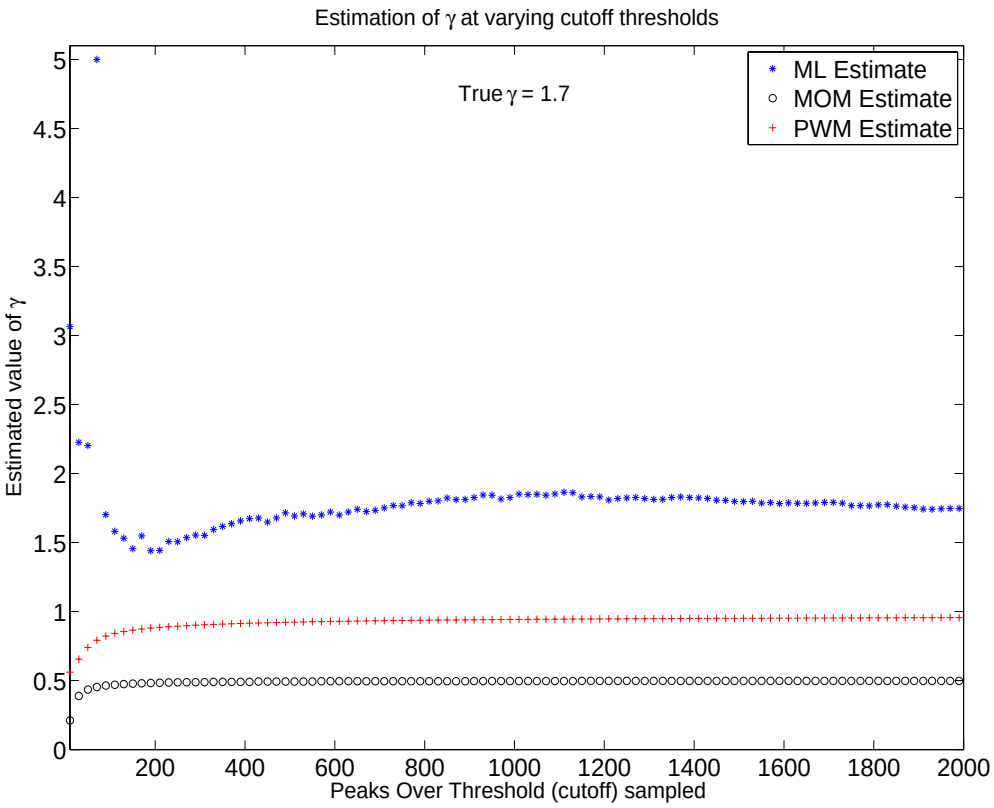

Estimation of γ at varying cutoff thresholds
True γ = 2.7
ML Estimate
MOM Estimate
PWM Estimate
Estimated value of γ
Peaks Over Threshold (cutoff) sampled
0
0.5
1
1.5
2
2.5
3
3.5
4
4.5
5
200
400
600
800
1000
1200
1400
1600
1800
2000

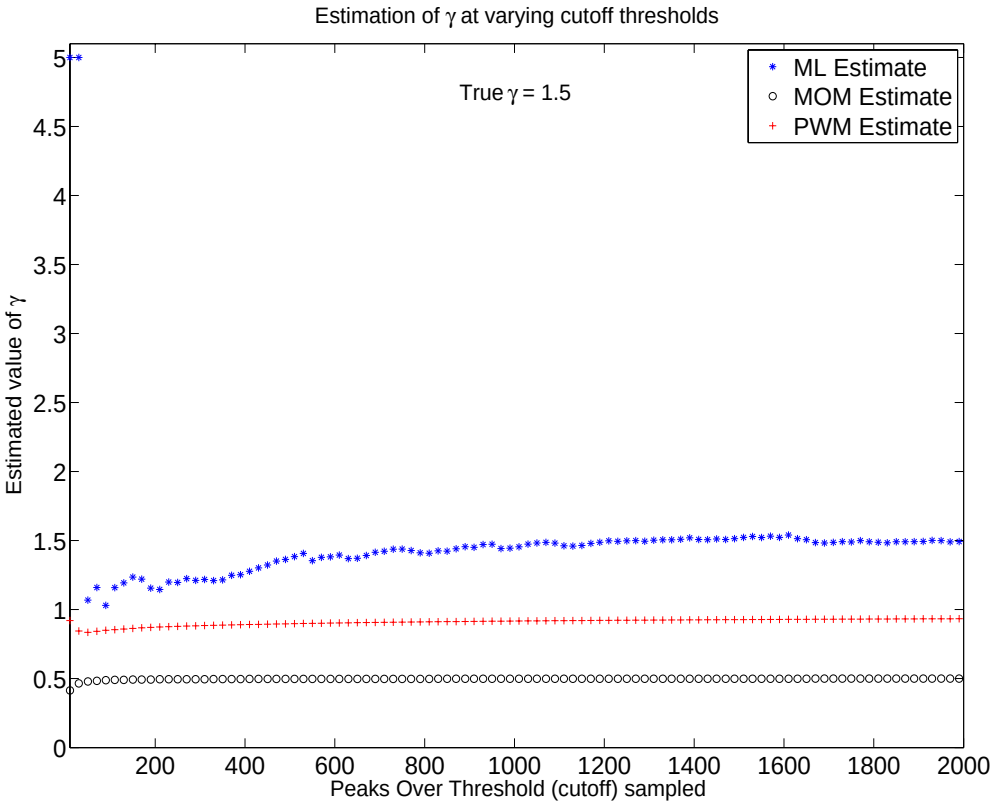

Estimation of γ at varying cutoff thresholds
True γ = 2.5
ML Estimate
MOM Estimate
PWM Estimate
Estimated value of γ
Peaks Over Threshold (cutoff) sampled
0
0.5
1
1.5
2
2.5
3
3.5
4
4.5
5
200
400
600
800
1000
1200
1400
1600
1800
2000

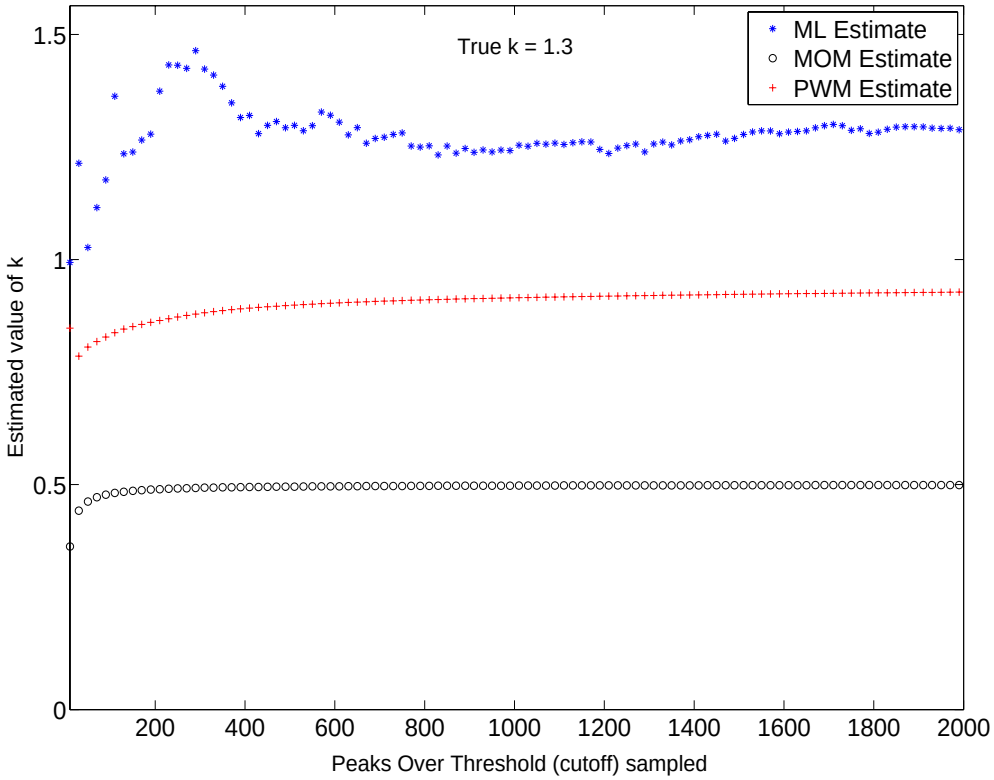

Estimation of γ at varying cutoff thresholds
True γ = 2.3
ML Estimate
MOM Estimate
PWM Estimate
Estimated value of γ
Peaks Over Threshold (cutoff) sampled
0
0.5
1
1.5
2
2.5
3
3.5
4
4.5
5
200
400
600
800
1000
1200
1400
1600
1800
2000

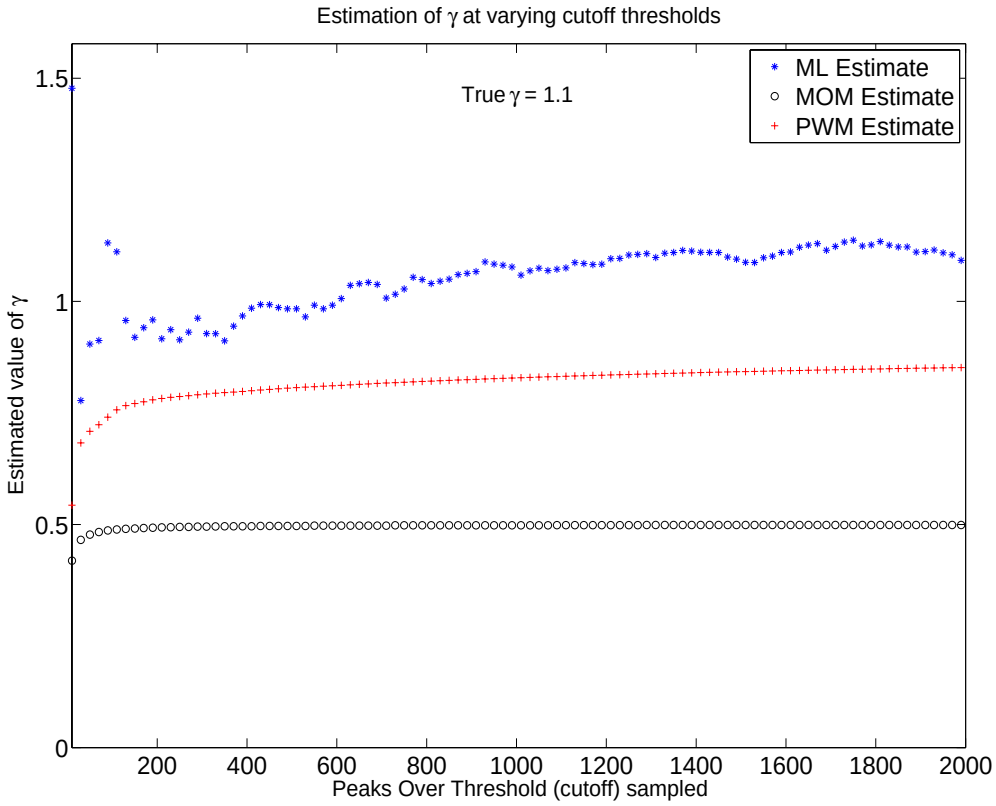

Estimation of γ at varying cutoff thresholds
True γ = 2.1
ML Estimate
MOM Estimate
PWM Estimate
Estimated value of γ
Peaks Over Threshold (cutoff) sampled
0
0.5
1
1.5
2
2.5
3
3.5
4
4.5
5
200
400
600
800
1000
1200
1400
1600
1800
2000

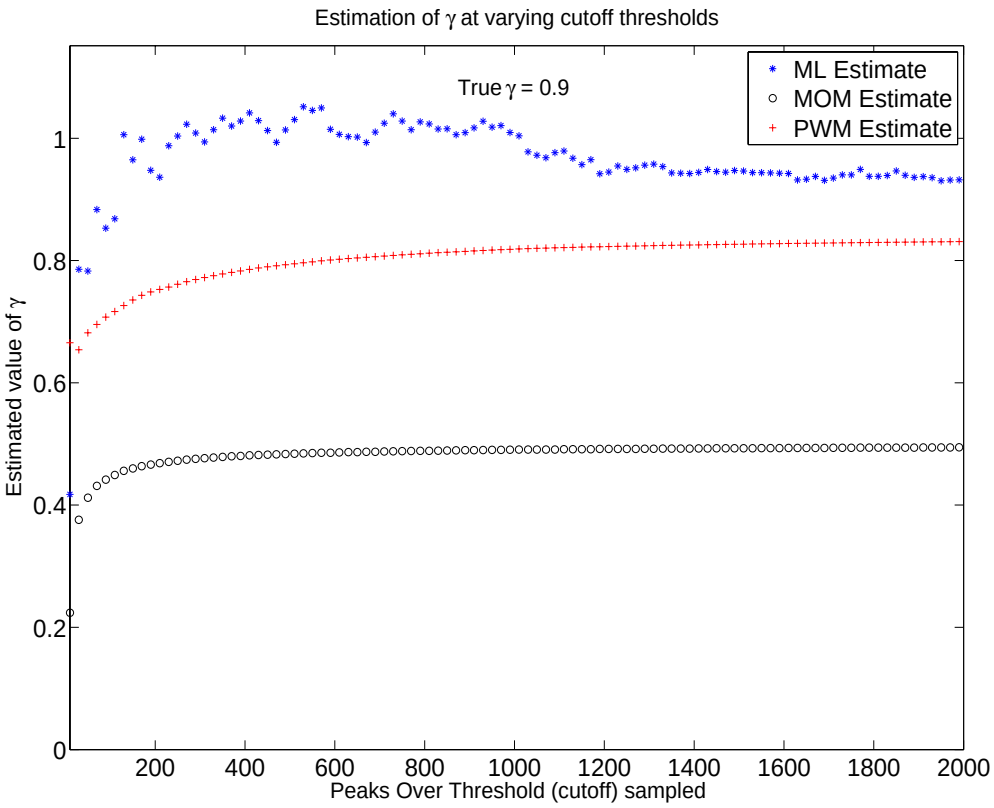

Estimation of γ at varying cutoff thresholds
True γ = 1.9
ML Estimate
MOM Estimate
PWM Estimate
Estimated value of γ
Peaks Over Threshold (cutoff) sampled
0
0.5
1
1.5
2
2.5
3
3.5
4
4.5
5
200
400
600
800
1000
1200
1400
1600
1800
2000

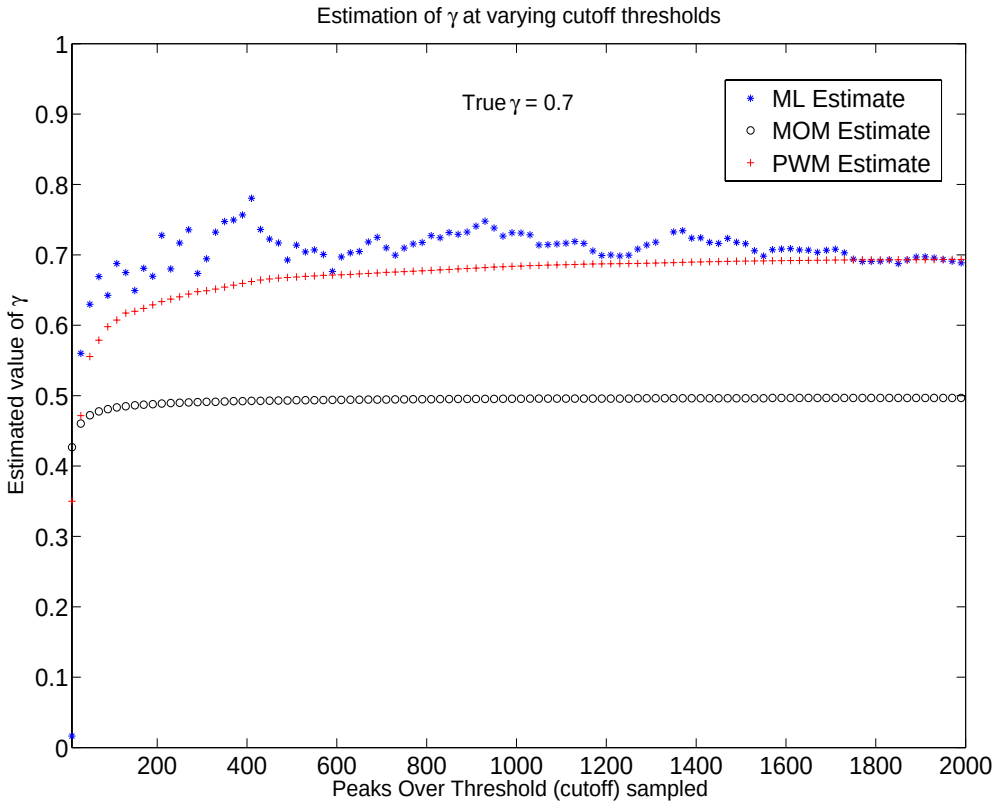

Estimation of γ at varying cutoff thresholds
True γ = 1.7
ML Estimate
MOM Estimate
PWM Estimate
Estimated value of γ
Peaks Over Threshold (cutoff) sampled
0
0.5
1
1.5
2
2.5
3
3.5
4
4.5
5
200
400
600
800
1000
1200
1400
1600
1800
2000

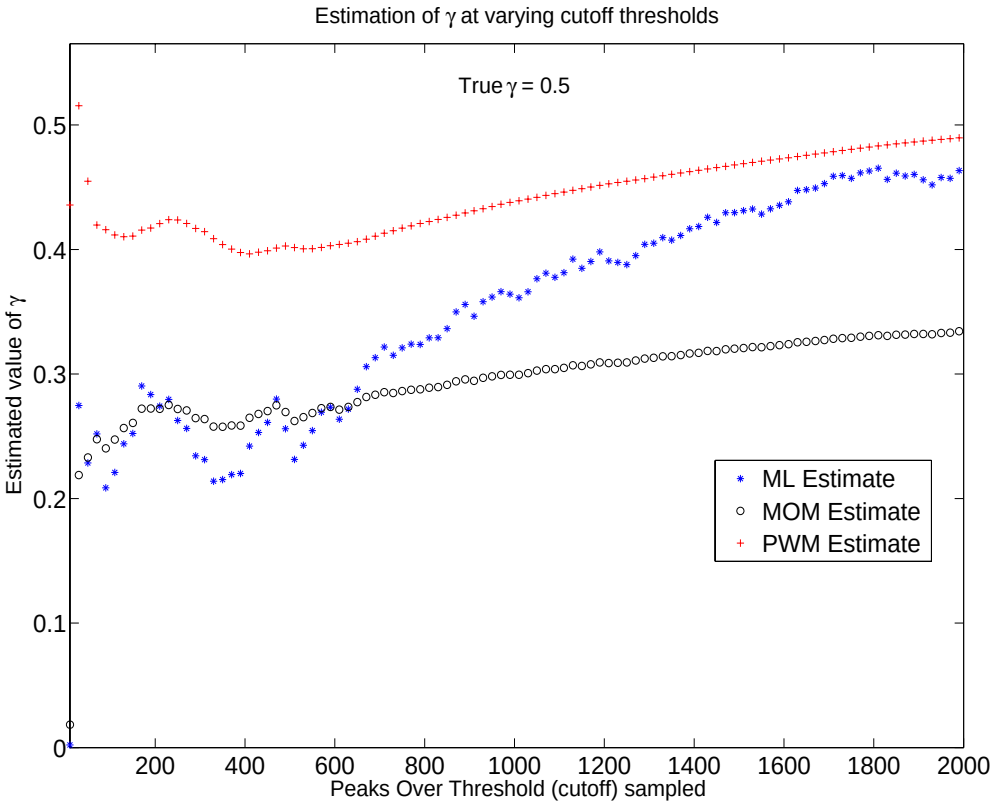

Estimation of γ at varying cutoff thresholds
True γ = 1.5
ML Estimate
MOM Estimate
PWM Estimate
Estimated value of γ
Peaks Over Threshold (cutoff) sampled
0
0.5
1
1.5
2
2.5
3
3.5
4
4.5
5
200
400
600
800
1000
1200
1400
1600
1800
2000

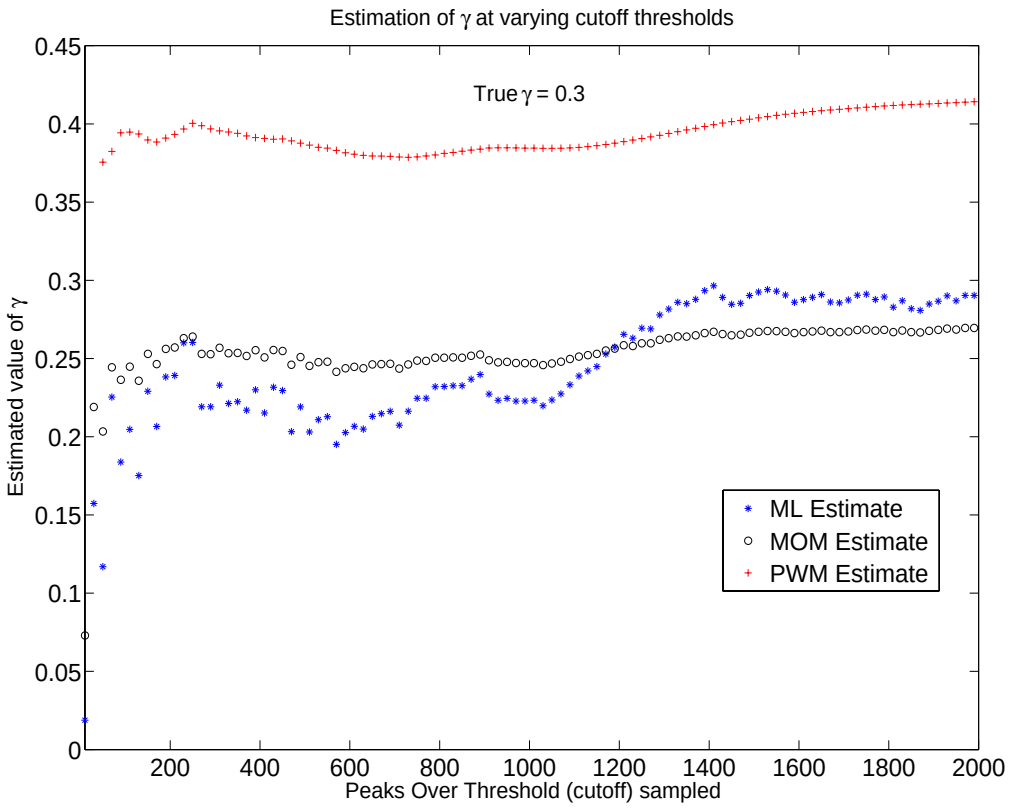

True k = 1.3
ML Estimate
MOM Estimate
PWM Estimate
Estimated value of k
Peaks Over Threshold (cutoff) sampled
1.5
1
0.5
0
200
400
600
800
1000
1200
1400
1600
1800
2000

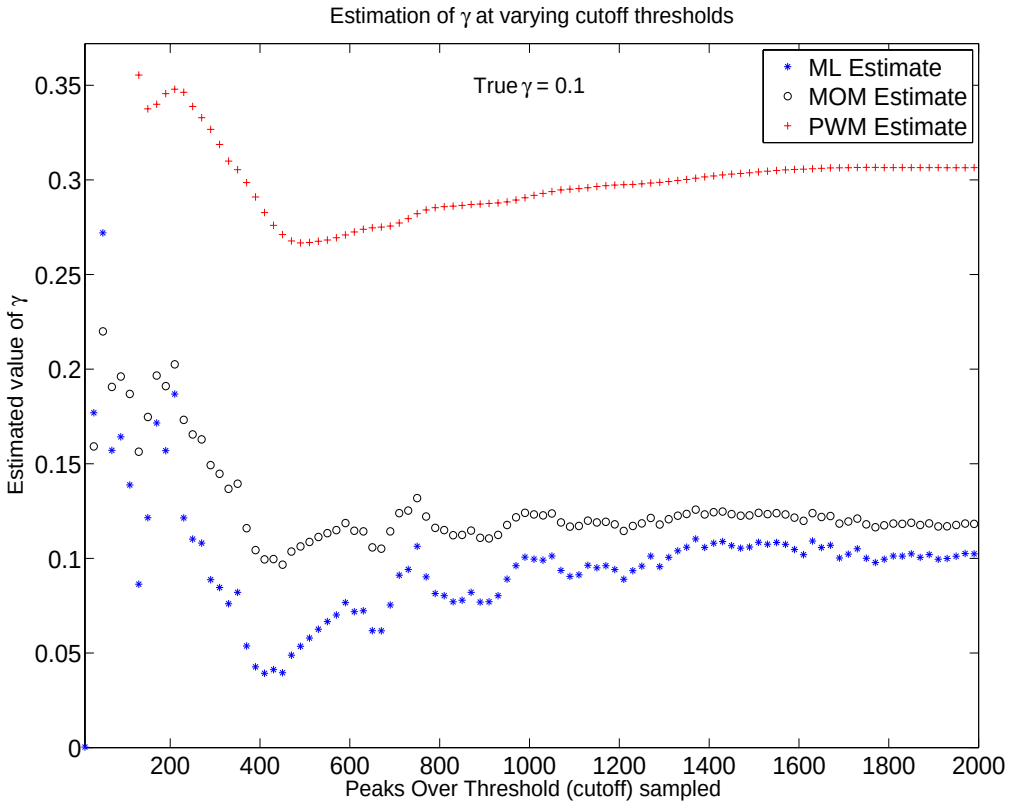

Estimation of γ at varying cutoff thresholds
True γ = 1.1
ML Estimate
MOM Estimate
PWM Estimate
Estimated value of γ
Peaks Over Threshold (cutoff) sampled
0
0.5
1
1.5
200
400
600
800
1000
1200
1400
1600
1800
2000


Estimation of γ at varying cutoff thresholds
True γ = 0.9
ML Estimate
MOM Estimate
PWM Estimate
Estimated value of γ
Peaks Over Threshold (cutoff) sampled
0
0.2
0.4
0.6
0.8
1
200
400
600
800
1000
1200
1400
1600
1800
2000


Estimation of γ at varying cutoff thresholds
True γ = 0.7
ML Estimate
MOM Estimate
PWM Estimate
Estimated value of γ
Peaks Over Threshold (cutoff) sampled
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1
200
400
600
800
1000
1200
1400
1600
1800
2000

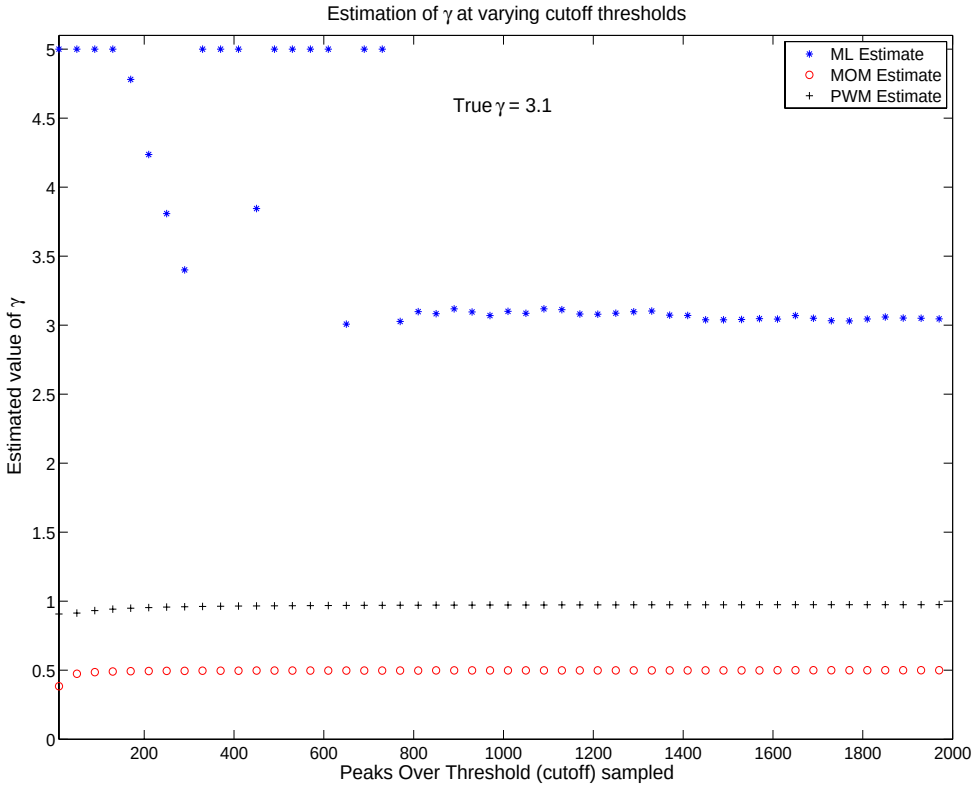

Estimation of γ at varying cutoff thresholds
True γ = 0.5
Estimated value of γ
Peaks Over Threshold (cutoff) sampled
ML Estimate
MOM Estimate
PWM Estimate
0
0.1
0.2
0.3
0.4
0.5
200
400
600
800
1000
1200
1400
1600
1800
2000

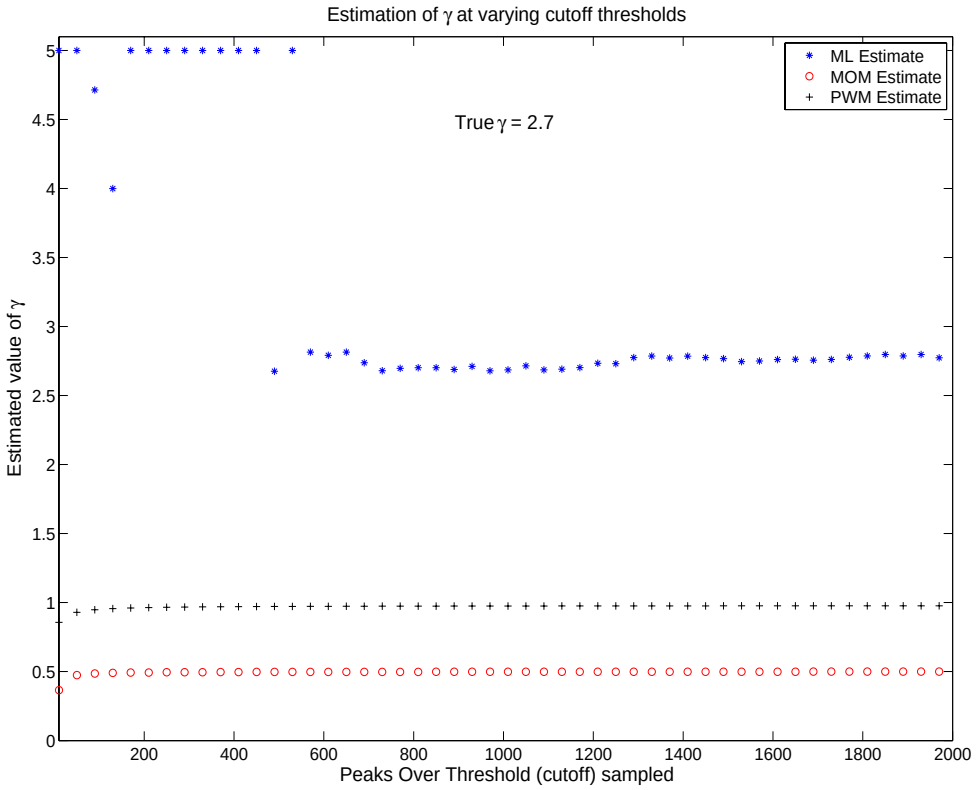

Estimation of γ at varying cutoff thresholds
True γ = 0.3
ML Estimate
MOM Estimate
PWM Estimate
Estimated value of γ
Peaks Over Threshold (cutoff) sampled
0.45
0.4
0.35
0.3
0.25
0.2
0.15
0.1
0.05
0
200
400
600
800
1000
1200
1400
1600
1800
2000

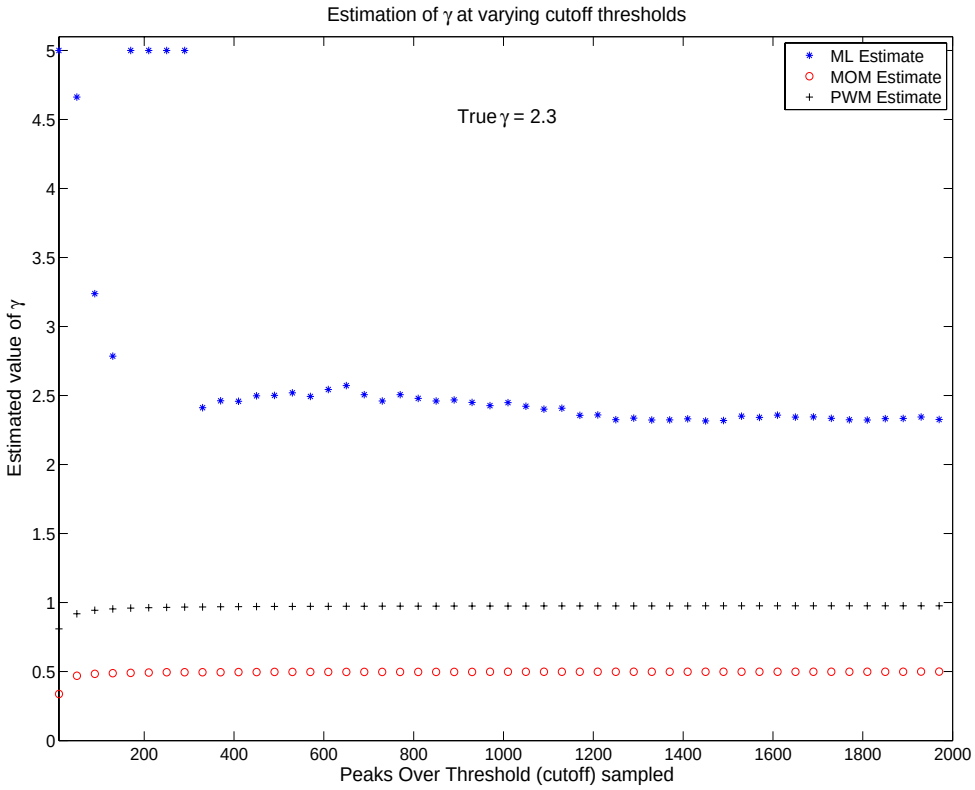

Estimation of γ at varying cutoff thresholds
True γ = 0.1
ML Estimate
MOM Estimate
PWM Estimate
Estimated value of γ
Peaks Over Threshold (cutoff) sampled
0
0.05
0.1
0.15
0.2
0.25
0.3
0.35
200
400
600
800
1000
1200
1400
1600
1800
2000

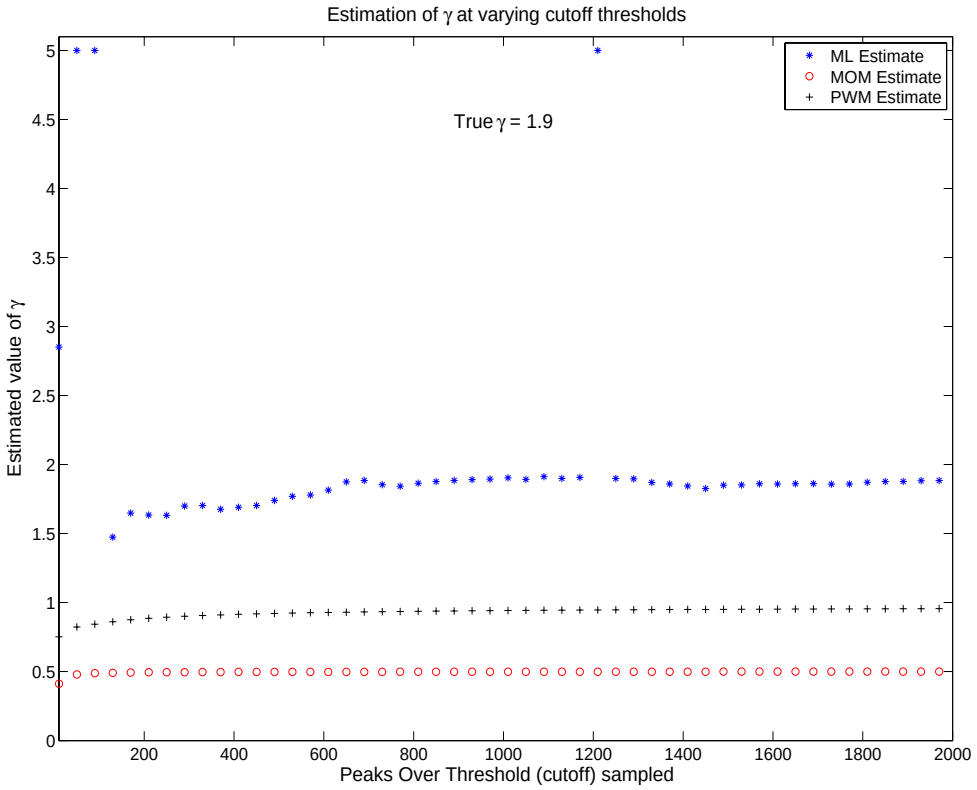
Appendix C. Simulations for ML, MOM and PWM Estimators

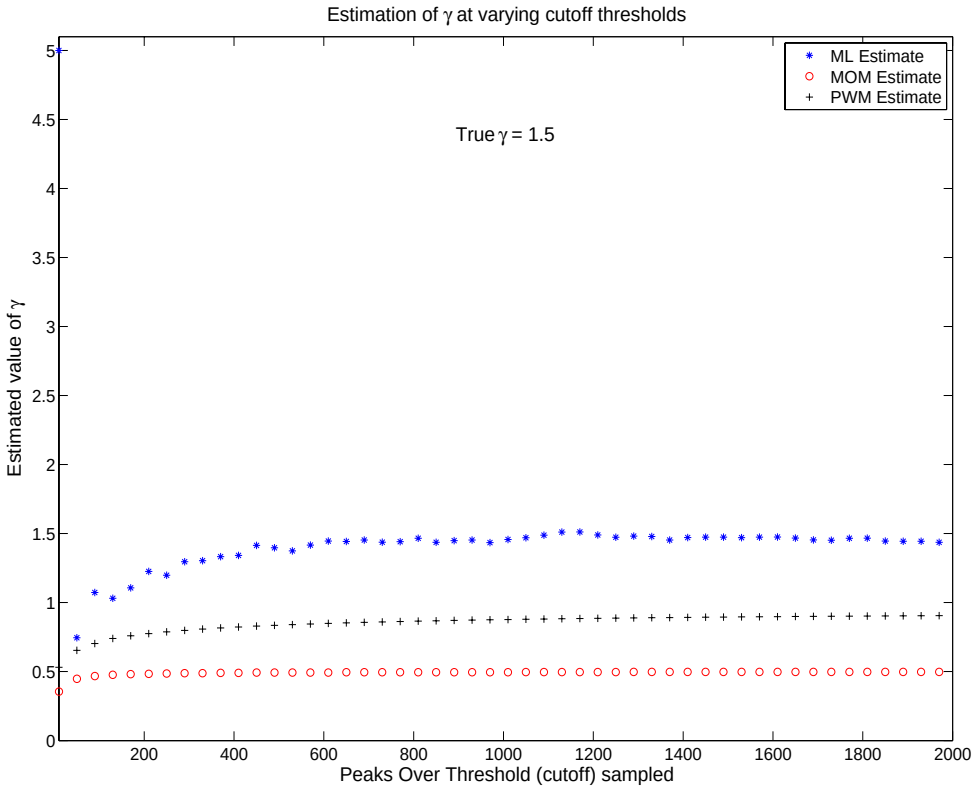
This appendix shows the results of ML, MOM and PWM estimators on simulated $|t_\nu|$ RVs:

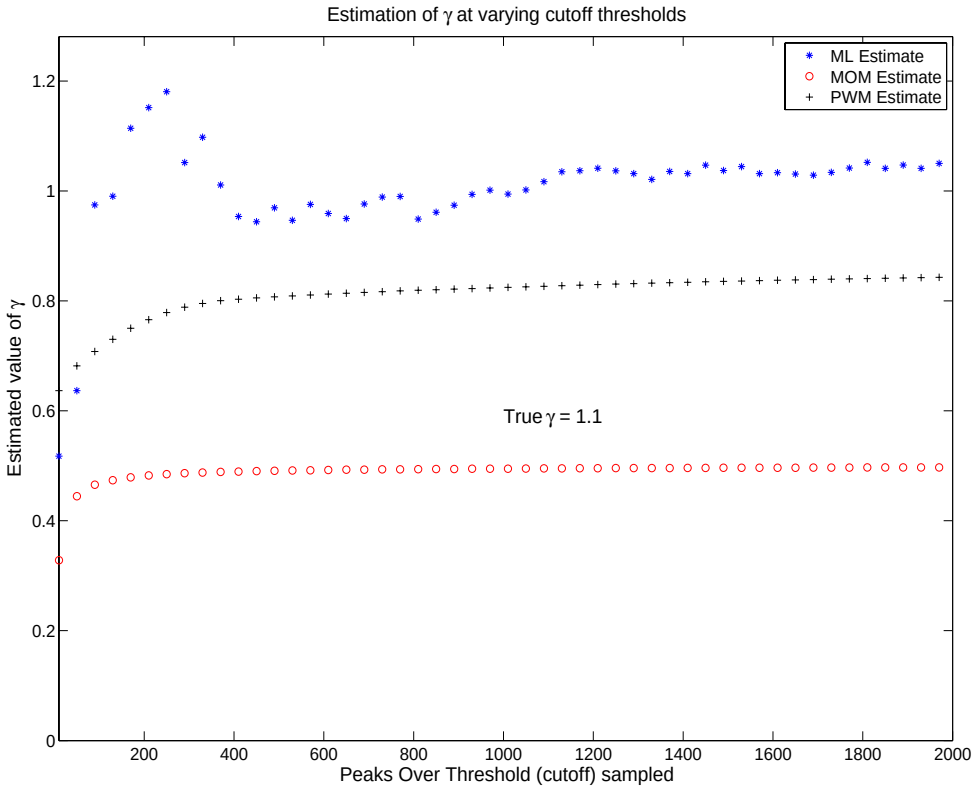


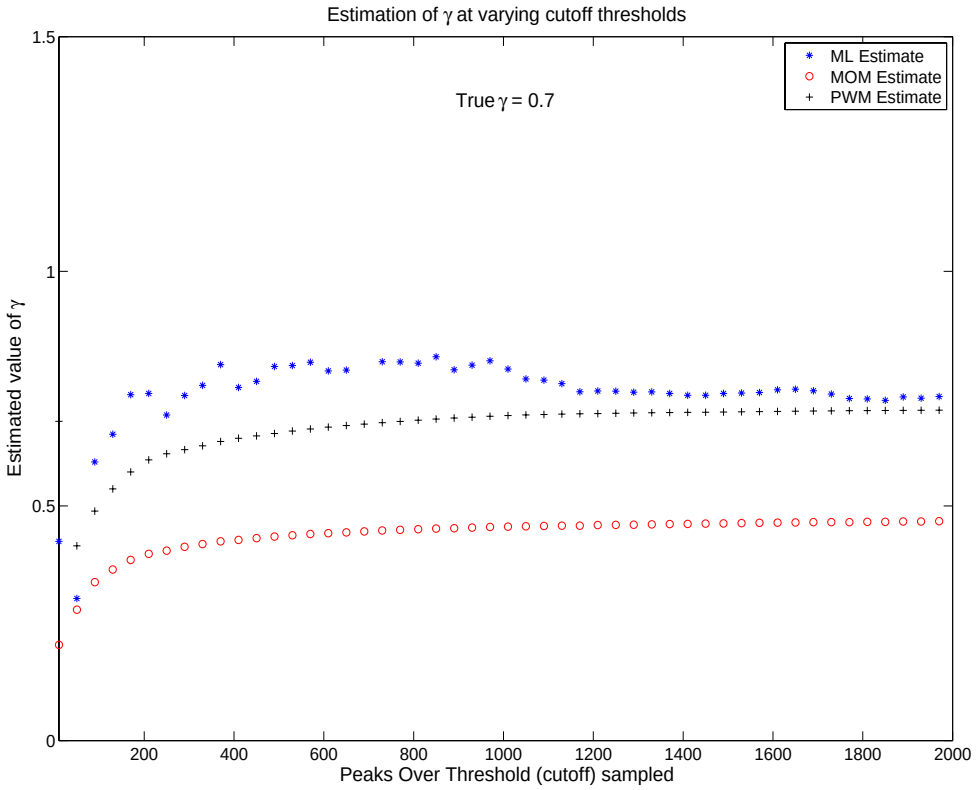

Estimation of γ at varying cutoff thresholds
ML Estimate
MOM Estimate
PWM Estimate
True γ = 2.7
Estimated value of γ
Peaks Over Threshold (cutoff) sampled
0
0.5
1
1.5
2
2.5
3
3.5
4
4.5
5
200
400
600
800
1000
1200
1400
1600
1800
2000

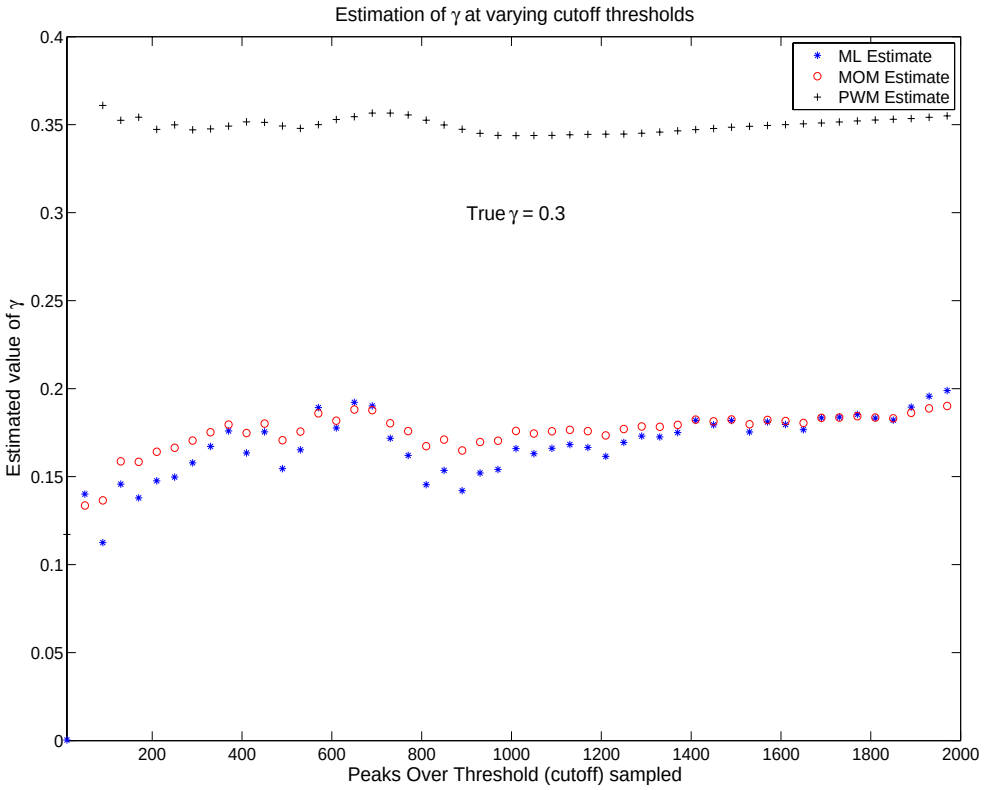

Estimation of γ at varying cutoff thresholds
ML Estimate
MOM Estimate
PWM Estimate
True γ = 2.3
Estimated value of γ
Peaks Over Threshold (cutoff) sampled
0
0.5
1
1.5
2
2.5
3
3.5
4
4.5
5
200
400
600
800
1000
1200
1400
1600
1800
2000

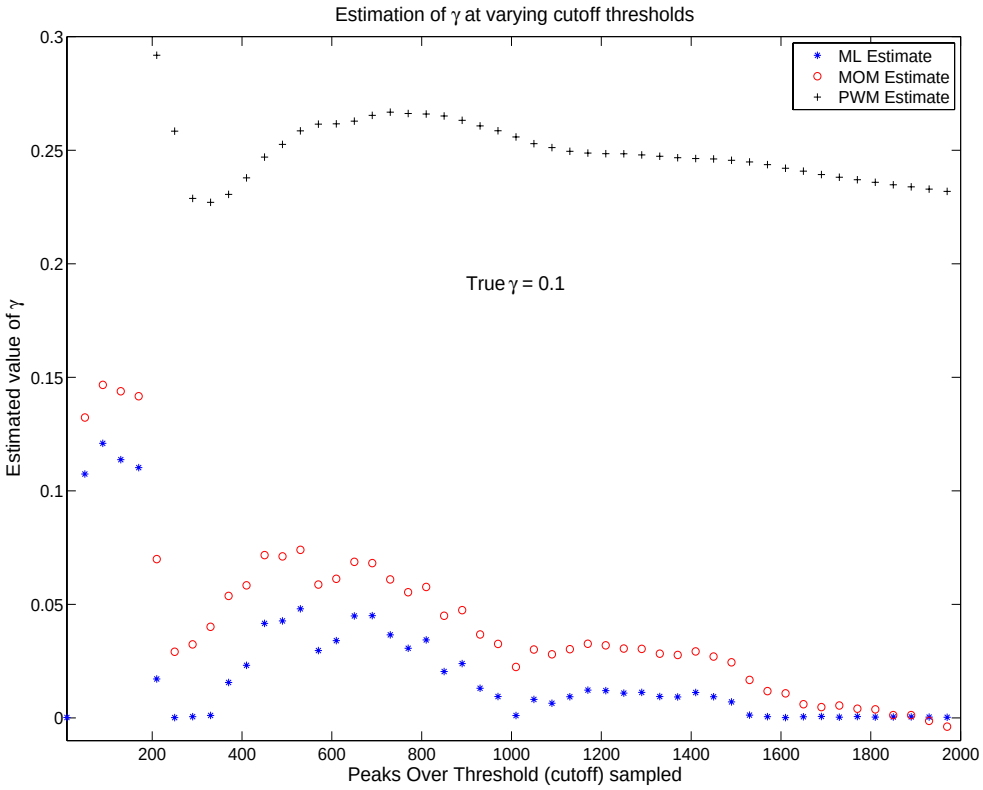

Estimation of γ at varying cutoff thresholds
ML Estimate
MOM Estimate
PWM Estimate
True γ = 1.9
Estimated value of γ
Peaks Over Threshold (cutoff) sampled
0
0.5
1
1.5
2
2.5
3
3.5
4
4.5
5
200
400
600
800
1000
1200
1400
1600
1800
2000


Estimation of γ at varying cutoff thresholds
ML Estimate
MOM Estimate
PWM Estimate
True γ = 1.5
Estimated value of γ
Peaks Over Threshold (cutoff) sampled
0
0.5
1
1.5
2
2.5
3
3.5
4
4.5
5
200
400
600
800
1000
1200
1400
1600
1800
2000


Estimation of γ at varying cutoff thresholds
ML Estimate
MOM Estimate
PWM Estimate
True γ = 1.1
Estimated value of γ
Peaks Over Threshold (cutoff) sampled
0
0.2
0.4
0.6
0.8
1
1.2
200
400
600
800
1000
1200
1400
1600
1800
2000


Estimation of γ at varying cutoff thresholds
ML Estimate
MOM Estimate
PWM Estimate
True γ = 0.7
Estimated value of γ
Peaks Over Threshold (cutoff) sampled
0
0.5
1
1.5
200
400
600
800
1000
1200
1400
1600
1800
2000


Estimation of γ at varying cutoff thresholds
ML Estimate
MOM Estimate
PWM Estimate
True γ = 0.3
Estimated value of γ
Peaks Over Threshold (cutoff) sampled
0
0.05
0.1
0.15
0.2
0.25
0.3
0.35
0.4
200
400
600
800
1000
1200
1400
1600
1800
2000


Estimation of γ at varying cutoff thresholds
ML Estimate
MOM Estimate
PWM Estimate
True γ = 0.1
Estimated value of γ
Peaks Over Threshold (cutoff) sampled
0.3
0.25
0.2
0.15
0.1
0.05
0
200
400
600
800
1000
1200
1400
1600
1800
2000

Appendix D. Simulations Results for EPM, ML, MOM and PWM Estimators

Tables of simulation results with GPD, t-distributed and F-distributed RVs.

Table D.1: Summary of Bias on Estimate of k using EPM Estimator on GPD data

cutoff	k = 0.1	k = 0.5	k = 1.0	k = 1.5	k = 2.0	k = 2.5
10	0.404	0.098	9.100	0.811	0.510	1.160
30	0.078	1.620	0.890	0.080	0.790	1.115
50	0.378	0.212	0.390	1.820	0.040	2.260
70	0.500	0.019	0.780	1.170	0.740	0.760
90	0.488	0.433	0.436	0.290	0.710	0.370
110	0.054	0.602	0.567	0.530	0.617	0.170
130	0.491	0.414	0.526	0.140	0.150	0.770
150	0.823	0.630	0.644	0.370	0.190	1.100
170	0.703	0.706	0.609	1.280	0.790	1.330
190	0.800	1.030	0.470	1.580	0.480	1.890
210	0.886	0.858	0.409	1.230	0.410	2.380
230	0.330	0.444	0.010	1.080	0.210	1.070
250	0.087	0.534	0.068	1.250	0.450	1.010
270	0.414	0.768	0.220	0.770	0.290	0.020
290	0.940	0.486	0.600	0.360	0.470	0.190
310	1.130	0.416	0.420	0.360	0.570	0.980
330	1.120	0.511	1.280	0.360	0.780	0.020
350	0.302	0.558	0.350	0.140	0.490	0.240
370	1.280	0.348	0.150	0.620	0.790	0.010
390	1.600	0.516	0.090	0.570	0.400	0.120
410	1.770	0.666	0.190	0.260	0.550	0.270
430	1.800	0.360	0.211	0.260	0.840	0.330
450	2.130	0.321	0.225	0.410	0.610	0.460
470	2.200	0.434	0.348	0.110	0.440	0.740
490	2.730	0.457	0.324	0.250	0.570	0.510
500	0.100	0.100	1.000	1.000	1.500	2.000

Table D.2: Summary of Bias on Estimate of k using MLE Estimator on GPD data

cutoff	k = 0.1	k = 0.5	k = 1.0	k = 1.5	k = 2.0	k = 2.5
10	1.200	0.820	0.330	0.130	0.530	0.430
30	0.213	0.151	0.690	0.280	0.060	0.060
50	0.299	0.022	0.670	0.450	0.160	0.410
70	0.249	0.172	0.560	0.370	0.120	0.340
90	0.213	0.026	0.750	0.180	0.150	0.290
110	0.129	0.055	0.760	0.180	0.410	0.270
130	0.146	0.036	0.680	0.040	0.230	0.350
150	0.172	0.057	0.700	0.120	0.130	0.380
170	0.173	0.061	0.630	0.220	0.030	0.370
190	0.159	0.086	0.540	0.240	0.060	0.380
210	0.148	0.075	0.480	0.210	0.050	0.380
230	0.124	0.030	0.370	0.180	0.070	0.190
250	0.087	0.037	0.340	0.170	0.030	0.160
270	0.107	0.060	0.290	0.110	0.040	0.050
290	0.116	0.025	0.250	0.030	0.010	0.060
310	0.121	0.010	0.230	0.010	0.000	0.150
330	0.123	0.016	0.170	0.010	0.030	0.010
350	0.119	0.017	0.190	0.040	0.000	0.010
370	0.108	0.013	0.190	0.000	0.030	0.030
390	0.109	0.006	0.190	0.020	0.010	0.060
410	0.109	0.004	0.180	0.060	0.010	0.090
430	0.108	0.025	0.220	0.070	0.050	0.110
450	0.112	0.037	0.220	0.050	0.020	0.130
470	0.108	0.020	0.230	0.100	0.010	0.160
490	0.114	0.019	0.230	0.080	0.010	0.150
500	0.114	0.018	0.220	0.078	0.010	0.140

Table D.3: Summary of Bias on Estimate of k using MOM Estimator on GPD data

cutoff	k = 0.1	k = 0.5	k = 1.0	k = 1.5	k = 2.0	k = 2.5
10	0.459	0.191	0.548	1.072	1.620	2.235
30	0.138	0.209	0.517	1.031	1.543	2.067
50	0.251	0.169	0.511	1.020	1.526	2.039
70	0.216	0.139	0.508	1.015	1.519	2.028
90	0.184	0.145	0.506	1.012	1.515	2.022
110	0.115	0.143	0.505	1.010	1.512	2.018
130	0.131	0.138	0.504	1.009	1.510	2.015
150	0.158	0.137	0.504	1.008	1.509	2.013
170	0.161	0.135	0.503	1.007	1.508	2.011
190	0.148	0.136	0.503	1.006	1.507	2.010
210	0.138	0.133	0.503	1.006	1.506	2.009
230	0.117	0.126	0.502	1.005	1.506	2.008
250	0.089	0.125	0.502	1.005	1.505	2.008
270	0.104	0.126	0.502	1.004	1.505	2.007
290	0.111	0.122	0.502	1.004	1.505	2.007
310	0.116	0.119	0.502	1.004	1.504	2.006
330	0.118	0.118	0.502	1.004	1.504	2.006
350	0.115	0.117	0.502	1.003	1.504	2.005
370	0.105	0.114	0.501	1.003	1.504	2.005
390	0.107	0.114	0.501	1.003	1.503	2.005
410	0.107	0.112	0.501	1.003	1.503	2.005
430	0.106	0.109	0.501	1.003	1.503	2.004
450	0.109	0.107	0.501	1.003	1.503	2.004
470	0.106	0.107	0.501	1.003	1.503	2.004
490	0.112	0.107	0.501	1.003	1.503	2.004
500	0.100	0.500	0.501	1.003	1.503	2.004

Table D.4: Summary of Bias on Estimate of k using PWM Estimator on GPD data

cutoff	k = 0.1	k = 0.5	k = 1.0	k = 1.5	k = 2.0	k = 2.5
10	0.331	0.192	0.530	0.733	1.187	1.911
30	0.157	0.067	0.409	0.811	1.161	1.646
50	0.106	0.066	0.308	0.797	1.127	1.597
70	0.110	0.027	0.259	0.784	1.107	1.577
90	0.094	0.020	0.227	0.766	1.094	1.565
110	0.122	0.026	0.204	0.749	1.085	1.558
130	0.129	0.028	0.187	0.735	1.077	1.553
150	0.129	0.027	0.173	0.721	1.072	1.549
170	0.118	0.025	0.162	0.712	1.067	1.546
190	0.115	0.026	0.154	0.705	1.064	1.543
210	0.115	0.026	0.147	0.699	1.061	1.541
230	0.120	0.023	0.142	0.693	1.058	1.540
250	0.128	0.018	0.137	0.687	1.056	1.538
270	0.138	0.017	0.134	0.682	1.054	1.537
290	0.140	0.014	0.131	0.678	1.052	1.536
310	0.142	0.011	0.129	0.673	1.051	1.535
330	0.141	0.007	0.127	0.669	1.049	1.535
350	0.140	0.004	0.125	0.665	1.048	1.534
370	0.140	0.001	0.124	0.661	1.047	1.533
390	0.142	0.002	0.122	0.657	1.046	1.533
410	0.141	0.005	0.121	0.654	1.045	1.532
430	0.141	0.008	0.120	0.651	1.044	1.532
450	0.142	0.011	0.119	0.648	1.044	1.531
470	0.142	0.014	0.117	0.645	1.043	1.531
490	0.141	0.017	0.116	0.642	1.042	1.531
500	0.140	0.016	0.115	0.642	1.042	1.531

Table D.5: Summary of Bias on Estimate of k using EPM Estimator on t-distributed data

cutoff	k = 0.1	k = 0.5	k = 1.0	k = 1.5	k = 2.0	k = 2.5
10	0.161	0.358	0.325	1.806	2.228	2.212
30	0.004	0.413	0.883	0.220	1.740	0.420
50	0.150	2.570	0.810	1.820	1.120	0.950
70	0.282	0.770	1.340	1.078	0.730	1.290
90	0.018	0.134	0.660	0.420	0.130	1.676
110	0.024	1.014	0.008	0.490	0.450	1.390
130	0.028	0.151	0.308	1.035	0.380	1.000
150	0.051	0.368	0.388	0.710	0.700	0.420
170	0.322	0.092	0.230	0.000	0.530	0.270
190	0.357	0.144	0.270	0.940	0.300	0.270
210	0.362	0.292	0.013	0.490	0.570	0.850
230	0.446	0.129	0.070	0.410	0.440	0.450
250	0.273	0.081	0.000	0.370	0.220	1.010
270	0.450	0.005	0.139	0.690	0.030	0.550
290	0.375	0.129	0.026	0.240	0.190	1.390
310	0.613	0.213	0.020	0.240	0.030	1.390
330	0.726	0.003	0.100	0.160	0.170	1.700
350	0.612	0.153	0.024	0.450	0.280	1.120
370	0.746	0.213	0.040	0.510	0.300	0.610
390	0.724	0.125	0.000	0.600	0.600	0.140
410	0.825	0.075	0.060	0.010	0.800	0.100
430	0.851	0.055	0.060	0.040	0.650	0.030
450	1.005	0.016	0.227	0.260	0.800	0.250
470	0.910	0.003	0.097	0.160	1.010	0.290
490	1.000	0.074	1.785	0.170	0.490	0.210
500	0.717	0.060	0.071	0.040	0.740	0.730

Table D.6: Summary of Bias on Estimate of k using MLE Estimator on t-distributed data

cutoff	k = 0.1	k = 0.5	k = 1.0	k = 1.5	k = 2.0	k = 2.5
10	1.180	0.776	0.390	0.100	0.350	0.010
30	0.400	0.505	0.600	0.230	0.220	0.750
50	0.220	0.468	0.070	0.320	0.540	0.490
70	0.188	0.402	0.040	0.240	0.430	0.050
90	0.084	0.278	0.031	0.190	0.410	0.050
110	0.035	0.254	0.030	0.200	0.330	0.210
130	0.052	0.178	0.090	0.260	0.300	0.340
150	0.078	0.264	0.100	0.180	0.220	0.440
170	0.075	0.206	0.010	0.080	0.230	0.520
190	0.113	0.199	0.004	0.180	0.240	0.550
210	0.142	0.196	0.020	0.090	0.190	0.400
230	0.137	0.170	0.010	0.070	0.190	0.570
250	0.124	0.125	0.020	0.060	0.210	0.610
270	0.114	0.129	0.050	0.090	0.220	0.570
290	0.100	0.136	0.020	0.030	0.230	0.600
310	0.115	0.136	0.020	0.020	0.200	0.580
330	0.111	0.104	0.010	0.010	0.170	0.580
350	0.114	0.120	0.020	0.040	0.140	0.520
370	0.132	0.123	0.010	0.050	0.130	0.450
390	0.142	0.108	0.020	0.050	0.080	0.360
410	0.144	0.095	0.010	0.030	0.050	0.340
430	0.151	0.086	0.030	0.030	0.060	0.310
450	0.150	0.076	0.060	0.010	0.040	0.310
470	0.143	0.069	0.030	0.010	0.010	0.260
490	0.142	0.054	0.060	0.000	0.070	0.250
500	0.127	0.070	0.030	0.030	0.040	0.180

Table D.7: Summary of Bias on Estimate of k using MOM Estimator on t-distributed data

cutoff	k = 0.1	k = 0.5	k = 1.0	k = 1.5	k = 2.0	k = 2.5
10	0.827	0.624	0.783	0.656	1.714	2.056
30	0.272	0.488	0.596	0.564	1.578	2.020
50	0.160	0.465	0.565	0.542	1.548	2.012
70	0.153	0.420	0.552	0.530	1.535	2.009
90	0.084	0.360	0.542	0.524	1.528	2.007
110	0.055	0.342	0.536	0.520	1.523	2.006
130	0.062	0.311	0.531	0.517	1.519	2.005
150	0.080	0.328	0.528	0.515	1.517	2.004
170	0.077	0.305	0.525	0.513	1.515	2.004
190	0.109	0.297	0.523	0.512	1.513	2.003
210	0.140	0.291	0.522	0.511	1.512	2.003
230	0.134	0.280	0.520	0.510	1.511	2.003
250	0.121	0.265	0.519	0.509	1.510	2.003
270	0.111	0.262	0.518	0.508	1.509	2.002
290	0.098	0.259	0.517	0.508	1.509	2.002
310	0.113	0.256	0.516	0.507	1.508	2.002
330	0.109	0.246	0.515	0.507	1.508	2.002
350	0.112	0.246	0.514	0.507	1.507	2.002
370	0.131	0.244	0.514	0.506	1.507	2.002
390	0.143	0.238	0.513	0.506	1.507	2.002
410	0.145	0.233	0.513	0.506	1.506	2.002
430	0.153	0.229	0.512	0.505	1.506	2.001
450	0.152	0.225	0.512	0.505	1.506	2.001
470	0.144	0.221	0.511	0.505	1.505	2.001
490	0.142	0.217	0.511	0.505	1.505	2.001
500	0.126	0.217	0.511	0.505	1.505	2.001

Table D.8: Summary of Bias on Estimate of k using PWM Estimator on t-distributed data

cutoff	k = 0.1	k = 0.5	k = 1.0	k = 1.5	k = 2.0	k = 2.5
10	0.379	0.336	0.484	0.722	1.320	1.868
30	0.099	0.203	0.267	0.730	1.227	1.704
50	0.114	0.214	0.225	0.715	1.178	1.670
70	0.144	0.202	0.224	0.696	1.159	1.643
90	0.165	0.191	0.220	0.677	1.144	1.623
110	0.192	0.160	0.215	0.661	1.132	1.609
130	0.199	0.132	0.210	0.649	1.123	1.599
150	0.184	0.119	0.205	0.639	1.116	1.591
170	0.179	0.118	0.201	0.631	1.109	1.585
190	0.167	0.111	0.199	0.625	1.104	1.580
210	0.154	0.104	0.197	0.619	1.099	1.576
230	0.139	0.099	0.195	0.615	1.095	1.572
250	0.132	0.091	0.193	0.611	1.091	1.569
270	0.128	0.084	0.191	0.607	1.088	1.566
290	0.130	0.079	0.189	0.604	1.085	1.564
310	0.131	0.074	0.187	0.601	1.083	1.562
330	0.131	0.069	0.186	0.599	1.080	1.560
350	0.129	0.066	0.184	0.596	1.078	1.559
370	0.128	0.063	0.183	0.594	1.076	1.557
390	0.123	0.060	0.182	0.592	1.074	1.556
410	0.119	0.056	0.180	0.590	1.073	1.554
430	0.115	0.053	0.179	0.589	1.071	1.553
450	0.111	0.049	0.178	0.587	1.070	1.552
470	0.109	0.045	0.177	0.586	1.068	1.551
490	0.107	0.042	0.176	0.585	1.067	1.550
500	0.107	0.039	0.175	0.583	1.066	1.549

Table D.9: Summary of Bias on Estimate of k using EPM Estimator on F-distributed data

cutoff	k = 0.1	k = 0.5	k = 1.0	k = 1.5	k = 2.0	k = 2.5
10	0.162	3.076	2.329	1.420	0.817	1.678
30	0.162	0.338	1.647	0.749	0.785	3.748
50	0.391	0.398	0.400	0.964	0.780	0.232
70	0.060	0.024	0.544	0.548	0.806	1.195
90	0.162	0.137	1.038	1.258	0.298	0.842
110	0.094	0.448	0.466	0.589	0.637	0.605
130	0.284	0.226	0.561	0.840	0.351	0.728
150	0.140	0.059	0.994	0.635	0.459	0.761
170	0.185	0.164	0.297	0.837	1.110	0.328
190	0.373	0.018	0.069	0.036	0.755	0.073
210	0.216	0.175	0.343	0.256	0.576	0.154
230	0.149	0.379	0.346	0.317	0.973	0.316
250	0.320	0.871	0.008	0.184	1.094	0.111
270	0.508	0.754	0.162	0.207	0.570	0.235
290	0.667	0.472	0.155	0.261	0.834	0.078
310	0.823	0.275	0.064	0.424	0.753	0.331
330	0.477	0.235	0.043	0.523	0.266	0.034
350	0.451	0.707	0.103	0.030	0.144	0.671
370	0.440	1.377	0.109	0.144	0.016	0.464
390	0.873	0.813	0.130	0.391	0.600	0.452
410	1.325	0.575	0.058	0.556	0.460	0.100
430	1.217	0.390	0.108	0.235	0.296	0.173
450	1.347	0.773	0.268	0.260	0.144	0.175
470	1.510	0.714	0.528	0.163	0.357	0.082
490	1.922	0.599	0.348	0.147	0.321	0.114
500	1.841	0.512	0.302	0.137	0.149	0.094

Table D.10: Summary of Bias on Estimate of k using MLE Estimator on F-distributed data

cutoff	k = 0.1	k = 0.5	k = 1.0	k = 1.5	k = 2.0	k = 2.5
10	0.079	0.632	0.091	0.361	0.059	0.105
30	0.087	0.119	0.078	0.267	0.602	0.203
50	0.088	0.183	0.013	0.441	0.383	0.483
70	0.147	0.034	0.041	0.364	0.558	0.608
90	0.129	0.009	0.166	0.367	0.370	0.289
110	0.154	0.088	0.190	0.241	0.477	0.388
130	0.110	0.076	0.205	0.212	0.294	0.366
150	0.140	0.039	0.235	0.136	0.373	0.535
170	0.132	0.009	0.173	0.110	0.420	0.432
190	0.097	0.025	0.148	0.047	0.376	0.368
210	0.127	0.043	0.156	0.037	0.332	0.381
230	0.136	0.075	0.150	0.050	0.329	0.391
250	0.104	0.117	0.111	0.122	0.310	0.306
270	0.074	0.116	0.075	0.136	0.232	0.351
290	0.060	0.106	0.064	0.151	0.234	0.283
310	0.068	0.092	0.065	0.054	0.200	0.248
330	0.076	0.082	0.072	0.050	0.119	0.287
350	0.076	0.115	0.053	0.133	0.088	0.213
370	0.075	0.146	0.047	0.114	0.059	0.236
390	0.057	0.124	0.041	0.081	0.102	0.240
410	0.022	0.108	0.058	0.067	0.076	0.288
430	0.018	0.091	0.061	0.105	0.048	0.275
450	0.026	0.115	0.080	0.110	0.020	0.328
470	0.024	0.112	0.105	0.126	0.033	0.283
490	0.016	0.105	0.090	0.131	0.021	0.273
500	0.015	0.102	0.094	0.131	0.026	0.356

Table D.11: Summary of Bias on Estimate of k using MOM Estimator on F-distributed data

cutoff	k = 0.1	k = 0.5	k = 1.0	k = 1.5	k = 2.0	k = 2.5
10	0.009	0.536	0.574	1.053	1.600	2.067
30	0.059	0.307	0.531	1.020	1.536	2.023
50	0.064	0.216	0.521	1.013	1.522	2.014
70	0.096	0.216	0.516	1.010	1.516	2.010
90	0.096	0.193	0.514	1.008	1.512	2.008
110	0.109	0.197	0.512	1.006	1.510	2.006
130	0.095	0.189	0.511	1.005	1.509	2.005
150	0.109	0.178	0.510	1.005	1.508	2.005
170	0.108	0.168	0.509	1.004	1.507	2.004
190	0.094	0.166	0.508	1.004	1.506	2.004
210	0.107	0.164	0.507	1.003	1.505	2.003
230	0.112	0.166	0.507	1.003	1.505	2.003
250	0.100	0.171	0.507	1.003	1.505	2.003
270	0.086	0.169	0.506	1.003	1.504	2.003
290	0.078	0.166	0.506	1.002	1.504	2.002
310	0.082	0.163	0.506	1.002	1.504	2.002
330	0.085	0.160	0.505	1.002	1.503	2.002
350	0.085	0.164	0.505	1.002	1.503	2.002
370	0.084	0.168	0.505	1.002	1.503	2.002
390	0.074	0.164	0.505	1.002	1.503	2.002
410	0.052	0.161	0.504	1.002	1.503	2.002
430	0.048	0.157	0.504	1.002	1.503	2.002
450	0.053	0.160	0.504	1.002	1.503	2.002
470	0.051	0.159	0.504	1.002	1.502	2.002
490	0.044	0.157	0.504	1.001	1.502	2.001
500	0.046	0.157	0.504	1.001	1.502	2.002

Table D.12: Summary of Bias on Estimate of k using PWM Estimator on F-distributed data

cutoff	k = 0.1	k = 0.5	k = 1.0	k = 1.5	k = 2.0	k = 2.5
10	0.455	0.025	0.591	0.880	1.060	1.641
30	0.337	0.122	0.414	1.011	1.083	1.594
50	0.237	0.024	0.403	0.945	1.087	1.573
70	0.253	0.018	0.368	0.899	1.083	1.561
90	0.257	0.006	0.359	0.866	1.080	1.554
110	0.264	0.005	0.362	0.842	1.076	1.549
130	0.262	0.015	0.361	0.820	1.073	1.545
150	0.261	0.016	0.361	0.801	1.070	1.542
170	0.263	0.010	0.357	0.784	1.068	1.540
190	0.255	0.005	0.352	0.770	1.066	1.538
210	0.256	0.002	0.347	0.756	1.064	1.537
230	0.259	0.005	0.343	0.745	1.062	1.536
250	0.257	0.009	0.338	0.735	1.061	1.535
270	0.253	0.015	0.333	0.725	1.059	1.534
290	0.246	0.018	0.328	0.717	1.058	1.533
310	0.240	0.019	0.323	0.709	1.057	1.532
330	0.238	0.019	0.319	0.703	1.056	1.532
350	0.236	0.020	0.315	0.697	1.055	1.531
370	0.235	0.024	0.311	0.691	1.054	1.531
390	0.232	0.027	0.307	0.686	1.053	1.530
410	0.227	0.027	0.303	0.681	1.052	1.530
430	0.219	0.027	0.300	0.677	1.051	1.530
450	0.215	0.026	0.298	0.673	1.050	1.529
470	0.211	0.027	0.295	0.669	1.049	1.529
490	0.207	0.027	0.293	0.665	1.049	1.529
500	0.201	0.026	0.29	0.661	1.049	1.529

Appendix E. Results from Threshold Sensitivity Analysis

Results from the threshold sensitivity analysis:

α = **1.3**, β = **1.0, for k = 0.1** u	RMSE of MAP	Mean 2nd derivative
10 - 100	0.21	0.06
100 - 200	0.21	0.00
200 - 300	0.17	0.00
300 - 400	0.15	0.00
400 - 500	0.13	0.00
500 - 900	0.10	0.00
α = **1.3**, β = **1.0, for k = 0.5** u	RMSE of MAP	Mean 2nd derivative
10 - 100	0.54	0.06
100 - 200	0.31	0.01
200 - 300	0.24	0.00
300 - 400	0.17	0.00
400 - 500	0.12	0.00
500 - 900	0.06	0.00
α = **1.3**, β = **1.0, for k = 1.0** u	RMSE of MAP	Mean 2nd derivative
10 - 100	1.33	0.09
100 - 200	0.74	0.02
200 - 300	0.50	0.01
300 - 400	0.35	0.01
400 - 500	0.25	0.00
500 - 900	0.11	0.00

Table E.1: MAP performance on F-distributed data with varying k. Gamma distributed prior used by Bayesian estimator (α and β parameter values shown))

α = **1.5**, β = **1.0, for k = 0.1**		
u	RMSE of MAP	Mean 2nd derivative
10 - 100	0.39	0.02
100 - 200	0.25	0.01
200 - 300	0.21	0.00
300 - 400	0.19	0.00
400 - 500	0.17	0.00
500 - 900	0.14	0.00
α = **1.5**, β = **1.0, for k = 0.5**		
u	RMSE of MAP	Mean 2nd derivative
10 - 100	0.71	0.06
100 - 200	0.42	0.01
200 - 300	0.30	0.01
300 - 400	0.23	0.00
400 - 500	0.18	0.00
500 - 900	0.11	0.00
α = **1.5**, β = **1.0, for k = 1.0**		
u	RMSE of MAP	Mean 2nd derivative
10 - 100	1.57	0.19
100 - 200	0.86	0.02
200 - 300	0.57	0.01
300 - 400	0.40	0.01
400 - 500	0.29	0.01
500 - 900	0.13	0.00

Table E.2: MAP performance on F-distributed data with varying k. Gamma distributed prior used by Bayesian estimator (α and β parameter values shown))

α = **1.9**, β = **1.0, for k = 0.1**		
u	RMSE of MAP	Mean 2nd derivative
10 - 100	0.16	0.03
100 - 200	0.18	0.00
200 - 300	0.19	0.00
300 - 400	0.18	0.00
400 - 500	0.16	0.00
500 - 900	0.13	0.00
α = **1.9**, β = **1.0, for k = 0.5**		
u	RMSE of MAP	Mean 2nd derivative
10 - 100	0.63	0.15
100 - 200	0.37	0.01
200 - 300	0.26	0.01
300 - 400	0.21	0.01
400 - 500	0.15	0.00
500 - 900	0.08	0.00
α = **1.9**, β = **1.0, for k = 1.0**		
u	RMSE of MAP	Mean 2nd derivative
10 - 100	1.42	0.17
100 - 200	0.74	0.03
200 - 300	0.48	0.02
300 - 400	0.32	0.01
400 - 500	0.23	0.01
500 - 900	0.10	0.00

Table E.3: MAP performance on F-distributed data with varying k. Gamma distributed prior used by Bayesian estimator (α and β parameter values shown))

$\alpha = 1.3$, $\beta = 1.0$, **for k = 0.1** u	RMSE of MAP	Mean 2nd derivative
10 - 100	0.19	0.01
100 - 200	0.18	0.01
200 - 300	0.16	0.00
300 - 400	0.14	0.00
400 - 500	0.12	0.00
500 - 900	0.09	0.00
$\alpha = 1.3$, $\beta = 1.0$, **for k = 0.5** u	RMSE of MAP	Mean 2nd derivative
10 - 100	0.44	0.05
100 - 200	0.27	0.01
200 - 300	0.20	0.00
300 - 400	0.14	0.00
400 - 500	0.10	0.00
500 - 900	0.06	0.00
$\alpha = 1.3$, $\beta = 1.0$, **for k = 1.0** u	RMSE of MAP	Mean 2nd derivative
10 - 100	1.29	0.07
100 - 200	0.70	0.02
200 - 300	0.44	0.01
300 - 400	0.25	0.01
400 - 500	0.23	0.00
500 - 900	0.10	0.00

Table E.4: MAP performance on GP-distributed data with varying k. Gamma distributed prior used by Bayesian estimator (α and β parameter values shown))

α = **1.5**, β = **1.0, for k = 0.1**		
u	RMSE of MAP	Mean 2nd derivative
10 - 100	0.38	0.04
100 - 200	0.25	0.01
200 - 300	0.19	0.00
300 - 400	0.16	0.00
400 - 500	0.15	0.00
500 - 900	0.12	0.00
α = **1.5**, β = **1.0, for k = 0.5**		
u	RMSE of MAP	Mean 2nd derivative
10 - 100	0.65	0.08
100 - 200	0.41	0.02
200 - 300	0.28	0.01
300 - 400	0.19	0.00
400 - 500	0.18	0.00
500 - 900	0.10	0.00
α = **1.5**, β = **1.0, for k = 1.0**		
u	RMSE of MAP	Mean 2nd derivative
10 - 100	1.38	0.15
100 - 200	0.80	0.03
200 - 300	0.50	0.02
300 - 400	0.38	0.01
400 - 500	0.31	0.00
500 - 900	0.15	0.00

Table E.5: MAP performance on GP-distributed data with varying k. Gamma distributed prior used by Bayesian estimator (α and β parameter values shown))

α = **1.9**, β = **1.0**, **for k = 0.1**		
u	RMSE of MAP	Mean 2nd derivative
10 - 100	0.39	0.05
100 - 200	0.25	0.01
200 - 300	0.19	0.00
300 - 400	0.19	0.00
400 - 500	0.17	0.00
500 - 900	0.13	0.00
α = **1.9**, β = **1.0**, **for k = 0.5**		
u	RMSE of MAP	Mean 2nd derivative
10 - 100	0.66	0.06
100 - 200	0.45	0.01
200 - 300	0.31	0.01
300 - 400	0.22	0.00
400 - 500	0.16	0.00
500 - 900	0.08	0.00
α = **1.9**, β = **1.0**, **for k = 1.0**		
u	RMSE of MAP	Mean 2nd derivative
10 - 100	1.45	0.13
100 - 200	0.90	0.04
200 - 300	0.58	0.02
300 - 400	0.39	0.01
400 - 500	0.25	0.00
500 - 900	0.12	0.00

Table E.6: MAP performance on GP-distributed data with varying k. Gamma distributed prior used by Bayesian estimator (α and β parameter values shown))

α = **1.3**, β = **1.0, for k = 0.1** u	RMSE of MAP	Mean 2nd derivative
10 - 100	0.21	0.02
100 - 200	0.20	0.01
200 - 300	0.17	0.00
300 - 400	0.14	0.00
400 - 500	0.13	0.00
500 - 900	0.10	0.00
α = **1.3**, β = **1.0, for k = 0.5** u	RMSE of MAP	Mean 2nd derivative
10 - 100	0.55	0.02
100 - 200	0.40	0.01
200 - 300	0.25	0.01
300 - 400	0.17	0.00
400 - 500	0.12	0.00
500 - 900	0.07	0.00
α = **1.3**, β = **1.0, for k =1.0** u	RMSE of MAP	Mean 2nd derivative
10 - 100	1.30	0.08
100 - 200	0.71	0.03
200 - 300	0.51	0.01
300 - 400	0.33	0.01
400 - 500	0.26	0.00
500 - 900	0.11	0.00

Table E.7: MAP performance on $|t_\nu|$-distributed data with varying k. Gamma distributed prior used by Bayesian estimator (α and β parameter values shown))

α = **1.5**, β = **1.0, for k = 0.1**		
u	RMSE of MAP	Mean 2nd derivative
10 - 100	0.41	0.02
100 - 200	0.28	0.01
200 - 300	0.21	0.01
300 - 400	0.17	0.00
400 - 500	0.16	0.00
500 - 900	0.14	0.00
α = **1.5**, β = **1.0, for k = 0.5**		
u	RMSE of MAP	Mean 2nd derivative
10 - 100	0.73	0.05
100 - 200	0.42	0.01
200 - 300	0.36	0.01
300 - 400	0.21	0.00
400 - 500	0.17	0.00
500 - 900	0.10	0.00
α = **1.5**, β = **1.0, for k = 1.0**		
u	RMSE of MAP	Mean 2nd derivative
10 - 100	1.52	0.17
100 - 200	0.91	0.05
200 - 300	0.70	0.01
300 - 400	0.38	0.01
400 - 500	0.31	0.00
500 - 900	0.15	0.00

Table E.8: MAP performance on $|t_\nu|$-distributed data with varying k. Gamma distributed prior used by Bayesian estimator (α and β parameter values shown))

α = **1.9**, β = **1.0, for k = 0.1** u	RMSE of MAP	Mean 2nd derivative
10 - 100	0.42	0.09
100 - 200	0.22	0.04
200 - 300	0.20	0.01
300 - 400	0.17	0.00
400 - 500	0.15	0.00
500 - 900	0.12	0.00
α = **1.9**, β = **1.0, for k = 0.5** u	RMSE of MAP	Mean 2nd derivative
10 - 100	0.77	0.09
100 - 200	0.52	0.02
200 - 300	0.38	0.01
300 - 400	0.24	0.00
400 - 500	0.16	0.00
500 - 900	0.08	0.00
α = **1.9**, β = **1.0, for k = 1.0** u	RMSE of MAP	Mean 2nd derivative
10 - 100	1.70	0.19
100 - 200	0.78	0.11
200 - 300	0.50	0.03
300 - 400	0.35	0.02
400 - 500	0.23	0.00
500 - 900	0.11	0.00

Table E.9: MAP performance on $|t_\nu|$-distributed data with varying k. Gamma distributed prior used by Bayesian estimator (α and β parameter values shown))

GPD with k = 0.1		
u	RMSE of ML	Mean 2nd derivative
10 - 100	0.20	0.10
100 - 200	0.08	0.02
200 - 300	0.10	0.01
300 - 400	0.10	0.01
400 - 500	0.04	0.01
500 - 900	0.02	0.01
GPD with k = 1.0		
u	RMSE of ML	Mean 2nd derivative
10 - 100	0.20	0.11
100 - 200	0.18	0.05
200 - 300	0.16	0.03
300 - 400	0.08	0.02
400 - 500	0.07	0.01
500 - 900	0.02	0.01
GPD with k = 2.0		
u	RMSE of ML	Mean 2nd derivative
10 - 100	0.28	0.30
100 - 200	0.24	0.07
200 - 300	0.19	0.04
300 - 400	0.15	0.03
400 - 500	0.04	0.01
500 - 900	0.03	0.01
GPD with k = 3.0		
u	RMSE of ML	Mean 2nd derivative
10 - 100	0.57	0.35
100 - 200	0.47	0.11
200 - 300	0.35	0.08
300 - 400	0.14	0.04
400 - 500	0.09	0.02
500 - 900	0.05	0.02

Table E.10: ML performance on GP-distributed data with varying k.

Average RMSE for $0.1 < k < 1.0$	
u	ARMSE
10-100	0.19
100-200	0.12
200-300	0.11
300-400	0.09
400-500	0.06
500-900	0.03
Average RMSE for $0.05 < k < 0.1$	
u	ARMSE
10-100	0.19
100-200	0.13
200-300	0.1
300-400	0.08
400-500	0.08
500-900	0.04
Average RMSE for $0.001 < k < 0.05$	
u	ARMSE
10-100	0.24
100-200	0.18
200-300	0.12
300-400	0.07
400-500	0.07
500-900	0.03
Average RMSE for $0.00001 < k < 0.001$	
u	ARMSE
10-100	0.3
100-200	0.22
200-300	0.13
300-400	0.09
400-500	0.05
500-900	0.02

Table E.11: Average RMSE of ML estimator for k values equally spaced within each of the limits: 1. $0.1 < k < 1.0$, 2. $0.05 < k < 0.1$, 3. $0.001 < k < 0.05$, 4. $0.00001 < k < 0.001$. Average values are taken from 10 GP distributed data sets with the k values equally spaced within each of the four regions.

Results from the ROI analysis using the methods developed in the research:

Figure F.1: The ROI of an assortment of various coniferous tree types. This ROI contains 23,411 pixels. It also contains many non-vegetative pixels, such as the pixels from the road at the top part of the ROI. The variability in this cluster of pixels is shown in Figure F.2. The MD data from this cluster are fit with a mixture of F-distributions, Johnson S_L distribution, and GPD with two optimized estimators for the tail-index parameter. The result is displayed in Figure F.3.

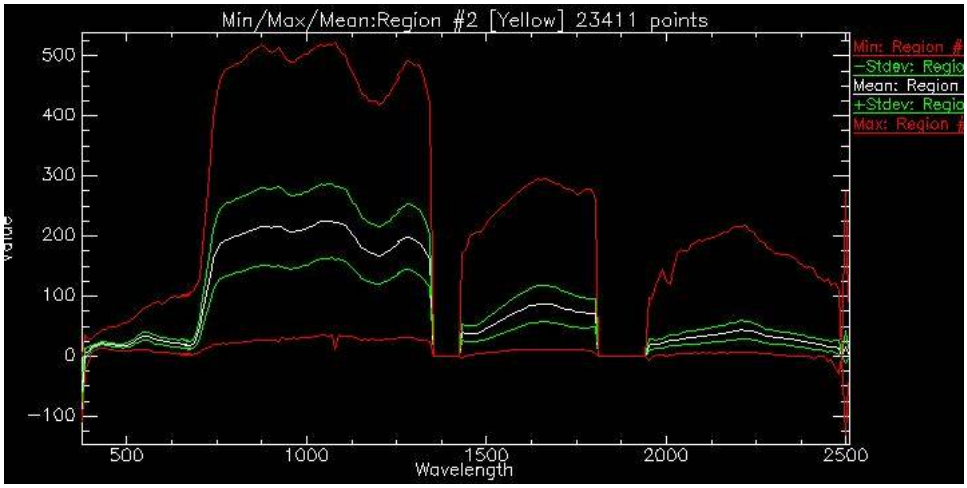


Figure F.2: The spectral variability for the ROI in Figure F.1. The y-axis values are units of spectral reflectance. The middle spectrum is the mean, the next two above and below it are the standard deviations, and the top and bottom spectra are the minimum and maximum in magnitude. Notice the increased variability in this ROI. This is due to the larger number of pixels encompassing more materials than just coniferous trees in the ROI.

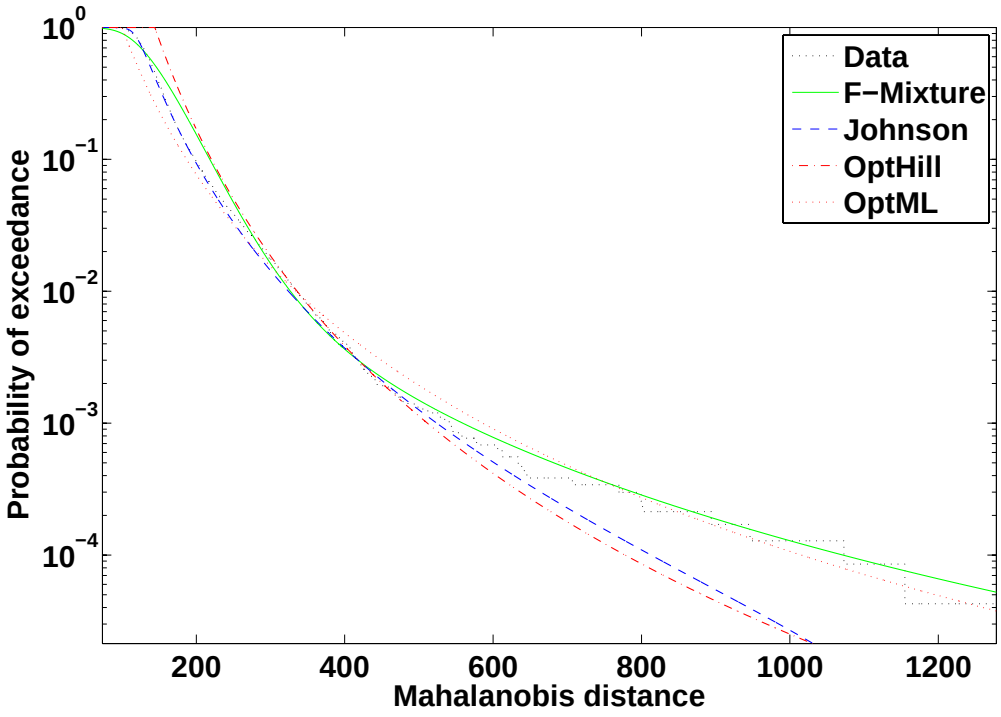


Figure F.3: Probability of exceedance plot showing the performance of each method used to fit the MD distribution from the MCFC ROI. Notice the larger MD values due to greater variability in the ROI.

Figure F.4: The ROI of an assortment of various deciduous tree types. This ROI contains 11,557 pixels. The variability in this cluster of pixels is shown in Figure F.5. The MD data from this cluster are fit with a mixture of F-distributions, Johnson S_L distribution, and GPD with two optimized estimators for the tail-index parameter. The result is displayed in Figure F.6.

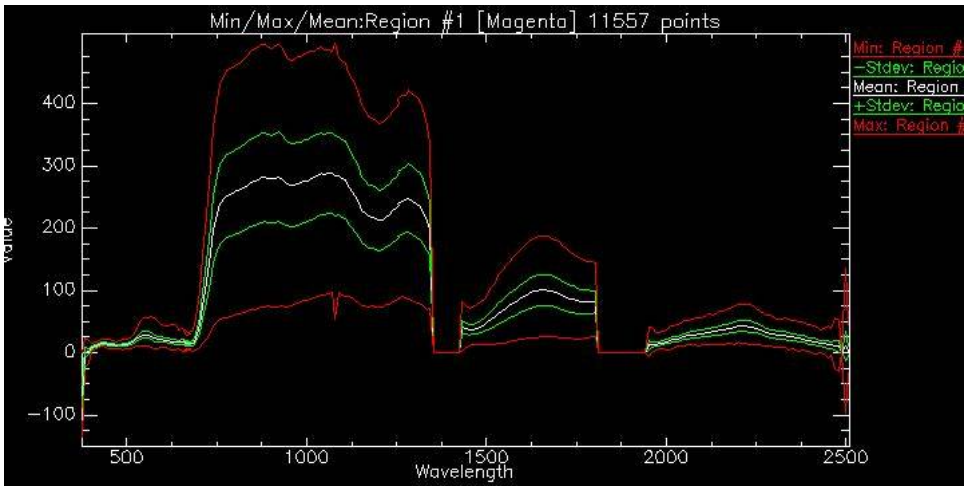


Figure F.5: The spectral variability for the ROI in Figure F.4. The y-axis values are units of spectral reflectance. The middle spectrum is the mean, the next two above and below it are the standard deviations, and the top and bottom spectra are the minimum and maximum in magnitude.

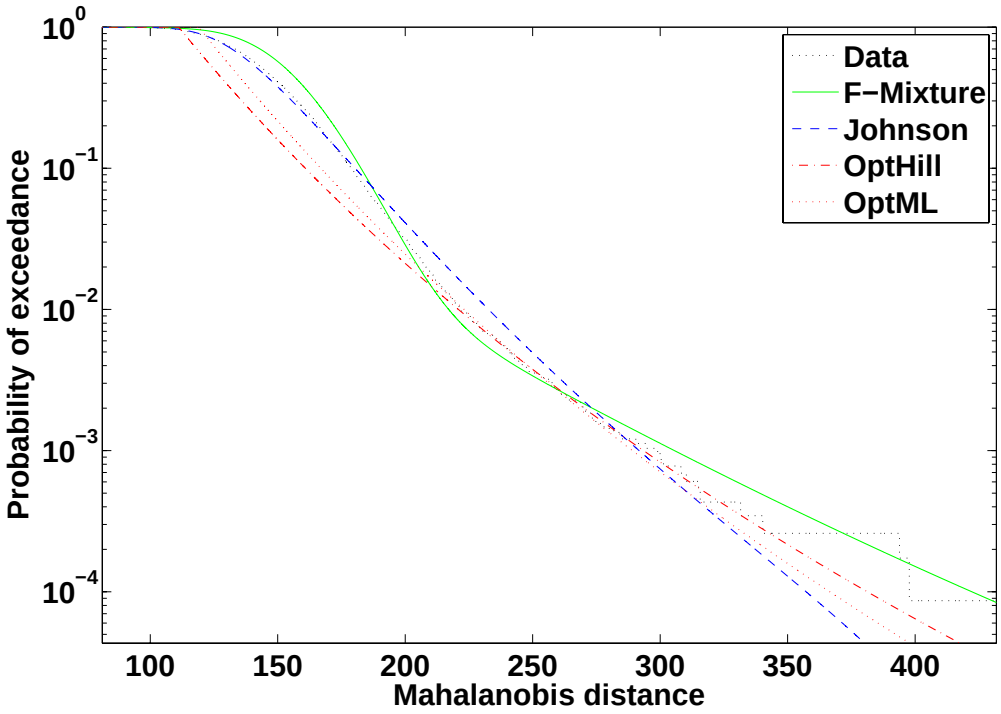


Figure F.6: Probability of exceedance plot showing the performance of each method used to fit the MD distribution from the DFC ROI. Notice how the F-mixture is affected by the last two data points (points most unlike the majority of the data). The optimized GPD methods and S_L will ignore those points.

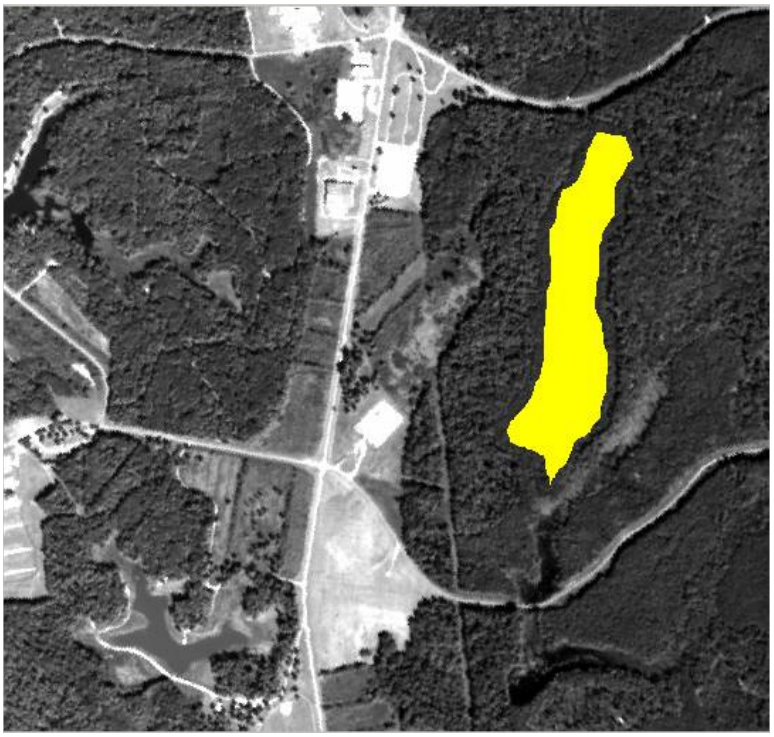


Figure F.7: The ROI of an assortment of various coniferous tree types. This ROI contains 9,212 pixels. The variability in this cluster of pixels is shown in Figure F.8. The MD data from this cluster are fit with a mixture of F-distributions, Johnson S_L distribution, and GPD with two optimized estimators for the tail-index parameter. The result is displayed in Figure F.9.

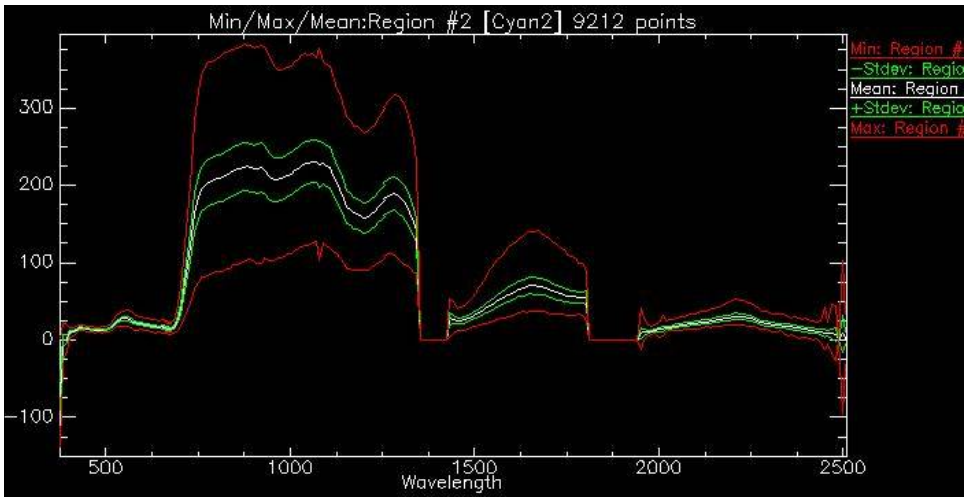


Figure F.8: The spectral variability for the ROI in Figure F.7. The y-axis values are units of spectral reflectance. The middle spectrum is the mean, the next two above and below it are the standard deviations, and the top and bottom spectra are the minimum and maximum in magnitude. Notice the decrease in variability compared to DFC and MCFC. For this ROI, the variability is purposely decreased by selecting a more homogenous coniferous forest area.

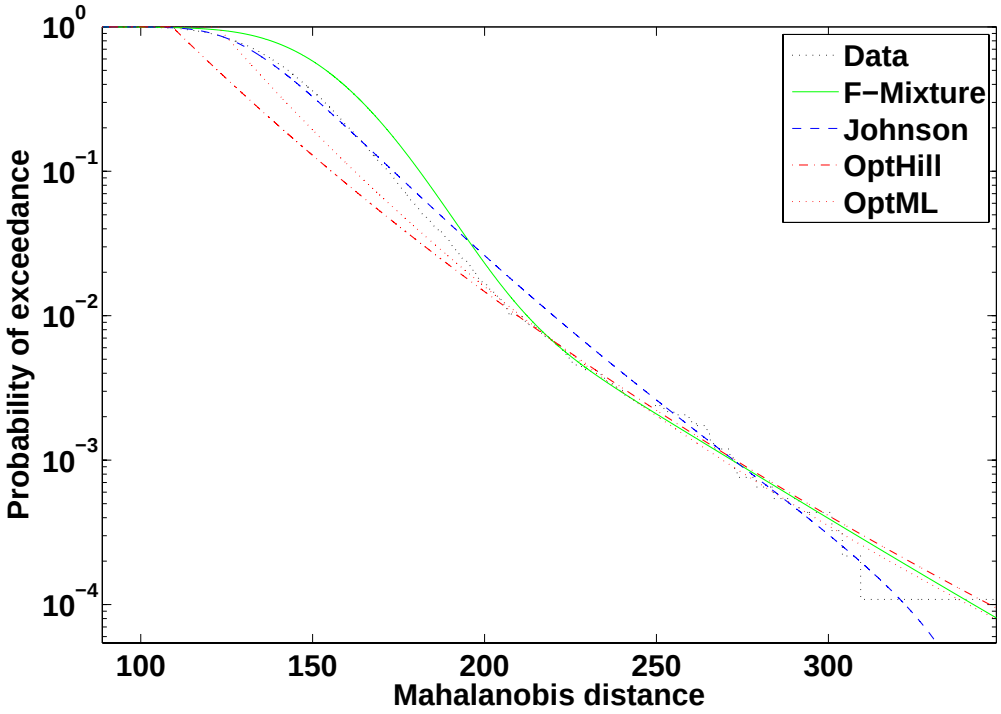


Figure F.9: Probability of exceedance plot showing the performance of each method used to fit the MD distribution from the CFC ROI. The MD values are smaller due to less variability in the ROI. Here, each fit is comparable. This should be the case, as this ROI contains a very homogeneous pixel set. However, checking the weighted MSE in Table 6.5, notice the disparity in fitting the end of the tail.

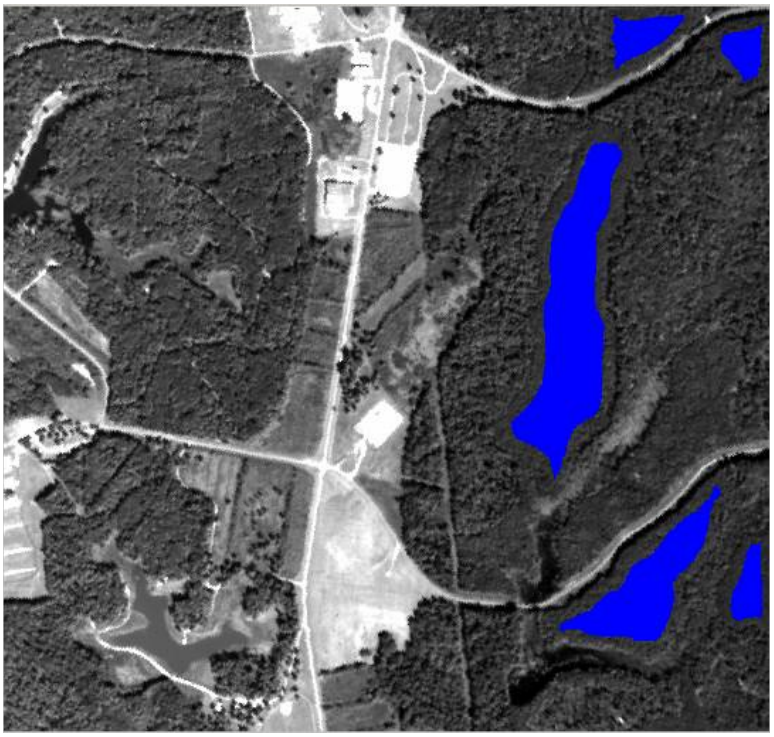


Figure F.10: The ROI of an assortment of Loblolly pine tree types. This ROI contains 14,257 pixels. The variability in this cluster of pixels is shown in Figure F.11. The MD data from this cluster are fit with a mixture of F-distributions, Johnson S_L distribution, and GPD with two optimized estimators for the tail-index parameter. The result is displayed in Figure F.12.

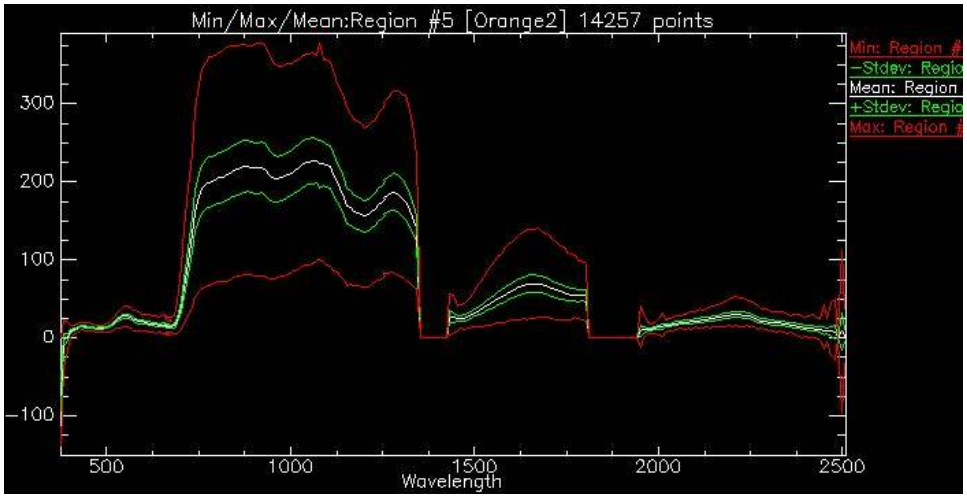


Figure F.11: The spectral variability for the ROI in Figure F.10. The y-axis values are units of spectral reflectance. The middle spectrum is the mean, the next two above and below it are the standard deviations, and the top and bottom spectra are the minimum and maximum in magnitude.

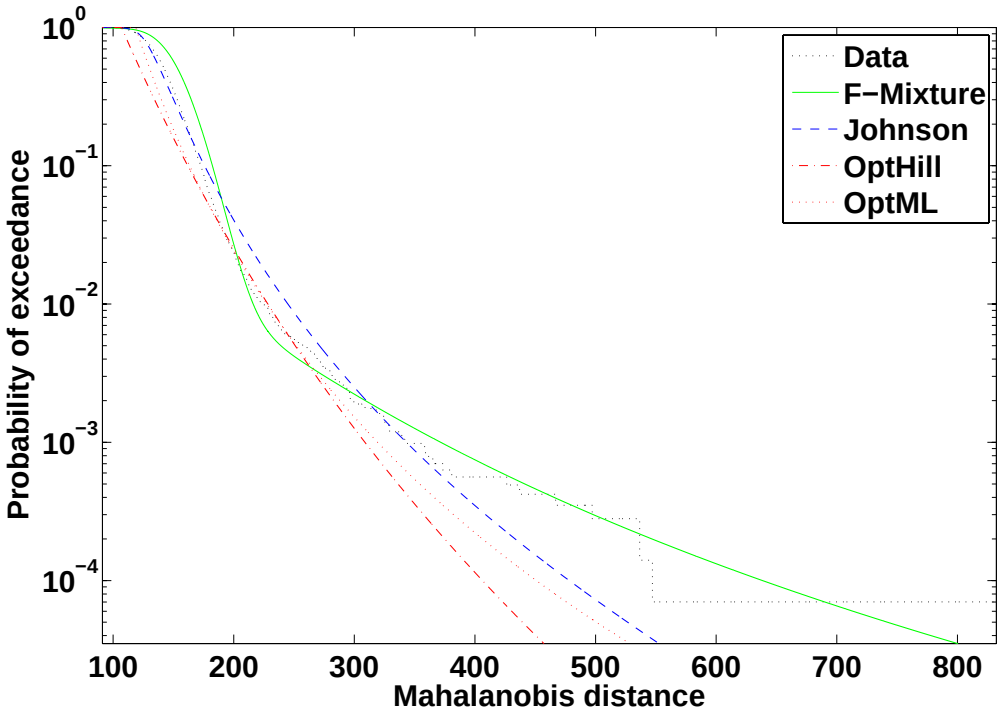


Figure F.12: Probability of exceedance plot showing the performance of each method used to fit the MD distribution from the ALPPC ROI. Again, the F-mixture tends to follow the largest extremes of the tail. In this case notice the "bump" in the region MD = 400 - 550. The optimized Hill, ML and S_L are developed to compensate for such a perturbation. This results in their optimal fit (the F-mixture overcompensates to fit the "bump" and final tail extremity, at the expense of worse performance in other regions of the tail).

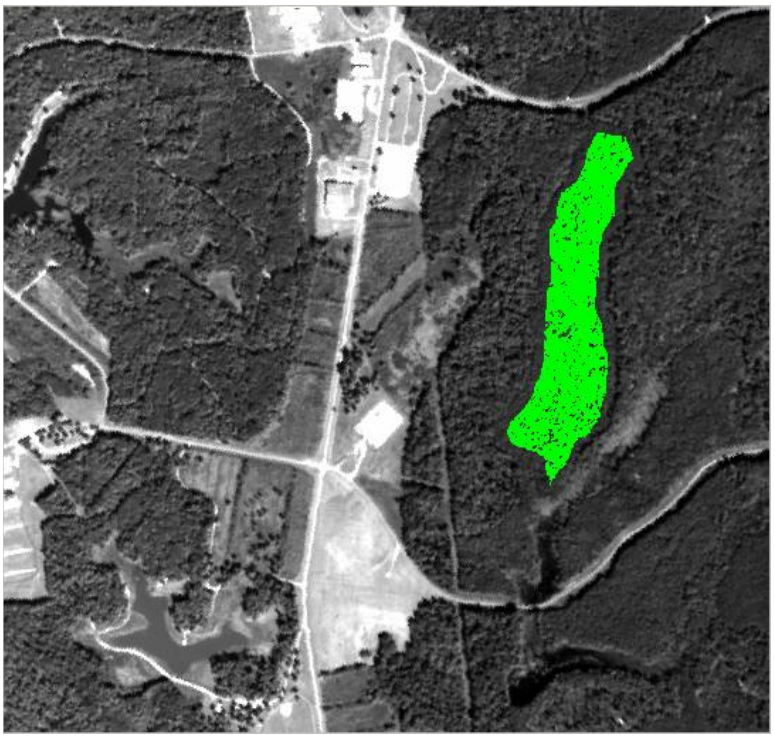


Figure F.13: The ROI of a reduced portion of CFC. This ROI contains 8,533 pixels. The ROI is reduced in size by eliminating pixels that decrease the homogeneity of the ROI material majority. The variability in this cluster of pixels is shown in Figure F.14. The MD data from this cluster are fit with a mixture of F-distributions, Johnson S_L distribution, and GPD with two optimized estimators for the tail-index parameter. The result is displayed in Figure F.15.

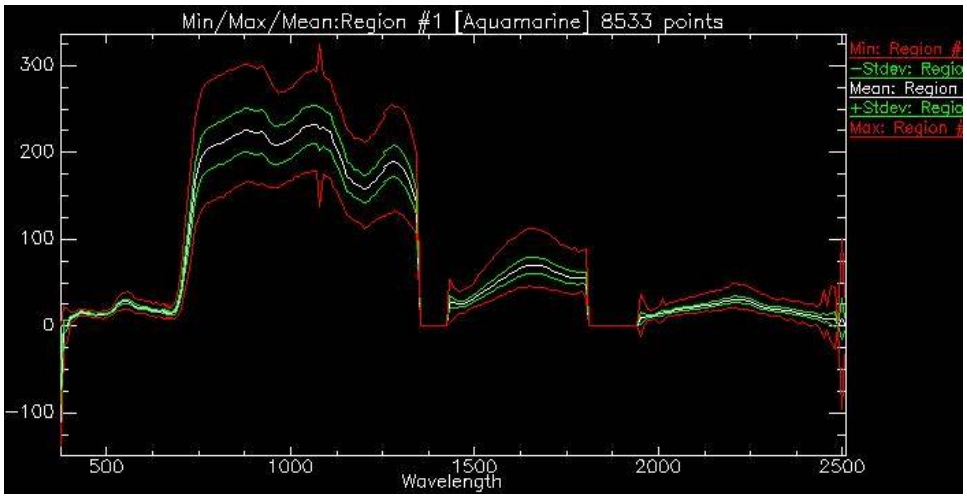


Figure F.14: The spectral variability for the ROI in Figure F.13. Notice the decreased variability due to the elimination of anomalous pixels.

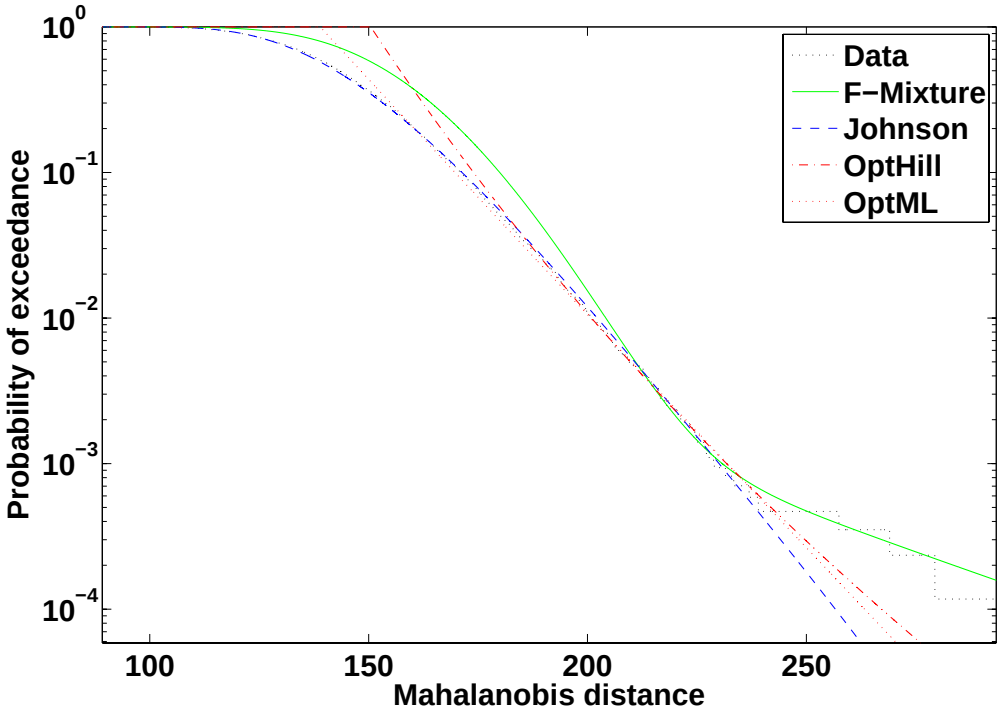


Figure F.15: Probability of exceedance plot showing the performance of each method used to fit the MD distribution from the CFCR ROI. The MD values are smaller compared to MD values from an ROI with greater variability. The "bump" starting at MD = 250 does not affect the S_L, optimized-Hill, and optimize-ML in this subset or in the same region found in Figure F.9.

Figure F.16: The ROI of a reduced portion of ALPPC. This ROI contains 12,976 pixels. The ROI is reduced in size by eliminating pixels that decrease the homogeneity of the ROI material majority. The variability in this cluster of pixels is shown in Figure F.17. The MD data from this cluster are fit with a mixture of F-distributions, Johnson S_L distribution, and GPD with two optimized estimators for the tail-index parameter. The result is displayed in Figure F.18.

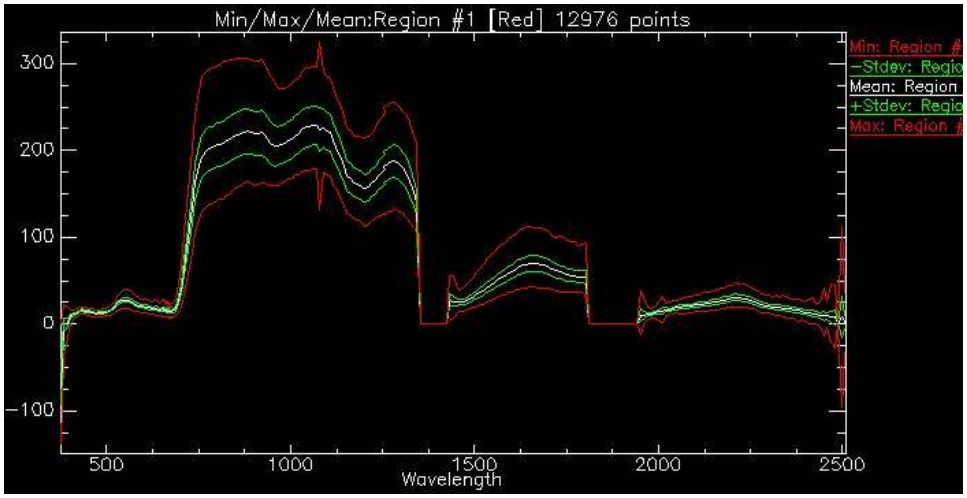


Figure F.17: The spectral variability for the ROI in Figure F.16. Notice the decreased variability compared to ALPPC due to the elimination of anomalous pixels.

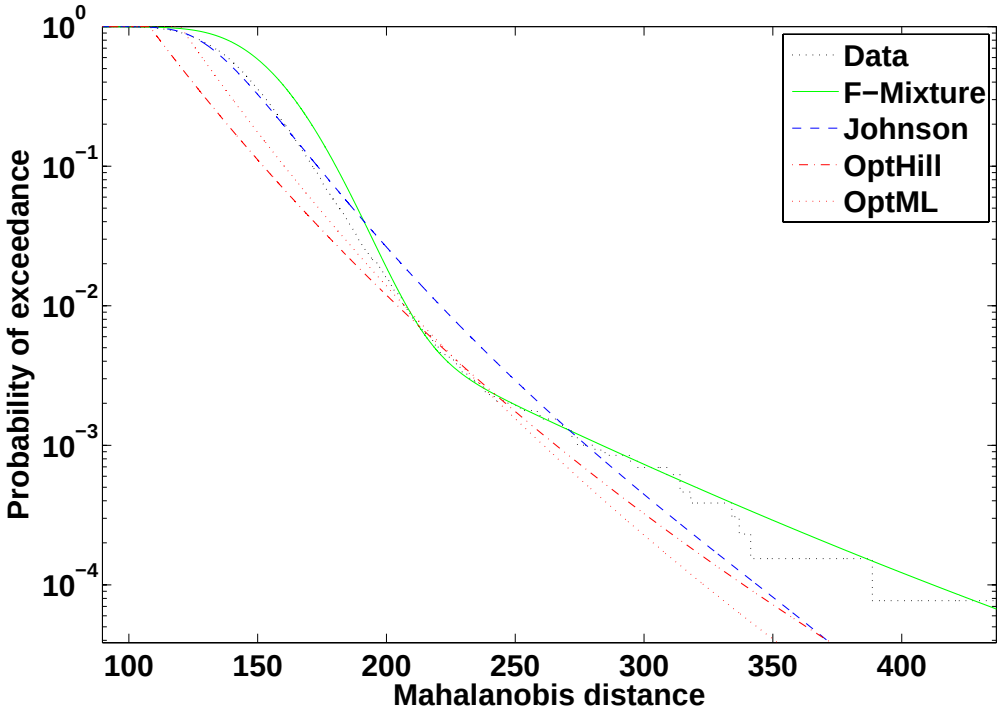


Figure F.18: Probability of exceedance plot showing the performance of each method used to fit the MD distribution from the ALPPCR ROI. The MD values are smaller compared to ALPPC MD values due to the decreased variability. This is due to the elimination of more anomalous pixels, leading to a reduction in the tail length (compare to Figure F.12) and improved MSE and weighted MSE (see Table 6.8 compared to Tabel 6.6.

Bibliography

1. *Projection-based adaptive anomaly detection for hyperspectral imagery*, volume 1, September 2003.

2. Anderson, T. W. *An Introduction to Multivariate Statistical Analysis*. John Wiley and Sons, Inc, 3 edition, 2003.

3. Beirlant, J., Y. Goegebeur, J. Segers, and J. Teugels. *Statistics of Extremes - Theory and Applications*. Wiley Series in Probability and Statistics, New York, 2005.

4. Berilant, J. and G. Matthys. *Adaptive Threshold Selection in Tail Index Estimation*. Catholic university press, Catholic University, University Center for Statistics, Leuven, Belgium, April 2000.

5. Berilant, J. and J. Teugels. *Practical Analysis of Extreme Values*. Technical report, Leuven University, Leuven, Belgium, 1996.

6. Boardman, J. "Analysis, Understanding and Visualization of Hyperspectral Data as Convex Sets in N-space". *SPIE Proceedings*, volume 2480, 14–22. 1995.

7. Box, G. E. P. and G. C. Tiao. *Bayesian Inference in Statistical Analysis*. Wiley, New York, 1992.

8. Cambanis, S., S. Huang, and G. Simons. "On the Theory of Elliptically Contoured Distributions". *Journal of Multivariate Analysis*, 11:638–385, 1981.

9. Castillo, E. *Extreme Value Theory in Engineering*. Academic Press, San Diego, 1988.

10. Castillo, E. and A. S. Hadi. "Fitting the Generalized Pareto Distribution to Data". *Journal of the American Statistical Association*, 1609–1621, December 1997.

11. Castillo, E., A. S. Hadi, N. Balakrishnan, and J. M. Sarabia. *Extreme Value and Related Models with Applications in Engineering and Science*. Wiley Series in Probability and Statistics, 2005.

12. Coles, S. *An introduction to Statistical Modeling of Extreme Values*. Springer Series in Statistics, London, 2001.

13. Coles, S. G. and E. A. Powell. "Bayesian Methods in Extreme Value Modelling: A Review and New Developments". *International Statistical Review*, 64:119–136, 1996.

14. Csorgo, S., P. DeHeuvels, and P. Mason. "Estimates of the Tail Index of a Distribution". *Annals of Statistics*, 13(3):1050 – 1077, 1985.

15. Dargahi-Noubary, G.R. "On Tail Estimation: An Improved Method". *International Association for Mathematical Geology*, 829 – 842, 1989.

16. Davison, A. C. and R. L. Smith. "Models for Exceedances Over High Thresholds". *Journal of the Royal Statistical Society, Series B*, 52(3):393–442, 1990.

17. Dekkers, A.L.M. and L. de Haan. "On the Estimation of the Extreme Value Index and Large Quantile Estimation". *Annals of Statistics*, 17:1833–1855, 1989.

18. Dempster, A., N. Laird, and D. Rubin. "Maximum Likelihood From Incomplete Data via the EM Algorithm". *Journal of the Royal Statistical Society. Series B (Methodological)*, 39(1):1–38, 1977.

19. Duda, Richard O., Peter E. Hart, and David G. Stork. *Pattern Classification*. John Wiley and Sons, Inc, 2 edition, 2001.

20. Embrechts, P., C. Kluppelberg, and T. Mikosch. *Modelling Ectremal Events for Insurance and Finance*. Springer, 2003.

21. Fang, K.T., S. Kotz, and K.W. Ng. *Symmetric Multivariate and Related Distributions*. Chapman and Hall, 1990.

22. Filzmoser, P., R.G. Garrett, and C. Riemann. "Multivariate Outlier Detection in Exploration Geochemistry". *Computers & Geosciences*, 31:Elsevier, 2005.

23. Gradshteyn, I. S. and I. M. Ryzhik. *Table of Integrals Series and Products, Sixth Ed.* Academic Press, 2000.

24. Green, R., G. Vane, T. Chrien, H. Enmark, E. Hansen, and W. Porter. "The Airborne Infrared Imaging Spectrometer". *Remote Sensing of the Environment*, 44:127–143, 1993.

25. Grimshaw, S.G. "Computing GPD Maximum Likelihood Estimates". *Technometrics*, 35(2):1993, 1993.

26. Groeneboom, P., H. P. Lopuhaä, and P. P. de Wolff. "Kernel-type Estimators for the Extreme Value Index". *Statistics*, August 2002.

27. Gruninger, J., A. J. Ratkowski, and M. L. Hoke. "The Sequential Maximum Angle Convex Cone (SMACC) Endmember Model". *Proceedings of SPIE*, volume 5425. 2004.

28. Hahn, G. J. and S. S. Shapiro. *Statistical Models in Engineering*. John Wiley & Sons, Inc., New York, 1967. Pp. 195-220.

29. Harsanyi, J.C. and C.-I. Chang. "Hyperspectral image classification and dimensionality reduction: an". *Geoscience and Remote Sensing, IEEE Transactions on*, 32(4):779–785, 1994.

30. Harsanyi, Joseph C. *Detection and Classification of Subpixel Spectral Signatures in Hyperspectral Image Sequences*. Ph.D. thesis, University of Maryland, 1993.

31. Hosking, J. R. M. and J. R. Wallis. "Parameter and Quantile Estimation for the Generalized Pareto Distribution". *Technometrics*, 29:339–349, 1987.

32. Hosking, J. R. M., J. R. Wallis, and E. F. Wood. "Estimation of the Generalized Extreme Value Distribution by the Method of Probability Weighted Moments". *Technometrics*, 27:251–261, 1985.

33. J. Theiler, B. R. Foy and A. M. Fraser. "Characterizing Non-Gaussian Clutter and Detecting Weak Gaseous Plumes in Hyperspectral Imagery". *Proceedings SPIE 5806*. 2005.

34. Johnson, M. E. *Multivariate Statistical Simulation*. John Wiley & Sons, Inc., 1967.

35. Johnson, N. L. "Systems of Frequency Curves Generated by Methods of Translation". *Biometrika*, 36:149–176, 1949.

36. Johnson, N. L., S. Kotz, and N. Balakrishnan. *Continuous Univariate Distributions, Vol 1., 2nd Edition*. Wiley Series in Probability and Mathematical Statistics, 1994.

37. Johnson, R. A. and D. W. Wichern. *Applied Multivariate Statistical Analysis*. Prentice Hall, 5 edition, 2002.

38. Kay, Steven M. *Fundamentals of Statistical Signal Processing, Vol I - Estimation Theory*, volume 1 of *Prentice Hall Signal Processing Series*. Prentice Hall, 1998.

39. Kay, Steven M. *Fundamentals of Statistical Signal Processing, Vol II - Detection Theory*, volume 2 of *Prentice Hall Signal Processing Series*. Prentice Hall, 7 edition, 1998.

40. Kelly, E.J. "An Adaptive Detection Algorithm". *IEEE Transactions on Aerospace and Electronic Systems*, 22(1):115–127, 1986.

41. Keshava, N., J. P. Kerekes, D. G. Manolakis, and G. A. Shaw. "Algorithm Taxonomy for Hyperspectral Unmixing". *Proceedings, SPIE*, volume 4049, 42–63. SPIE, 2000.

42. Kotz, S., T. J. Kozubowski, and K. Podgorski. "An Asymmetric Multivariate Laplae Distribution", January 2003. Department of Engineering, Management & System Engineering, The George Washington University, Washington, DC 20052.

43. Kraut, S., L. L. Scharf, and T. L. McWhorter. "Adaptive Subspace Detectors". *IEEE Transactions on Signal Processing*, 29(1):1–16, January 2001.

44. Kraut, S. and L.L. Scharf. "The CFAR adaptive subspace detector is a scale-invariant GLRT". *Signal Processing, IEEE Transactions on [see also Acoustics, Speech, and Signal Processing, IEEE Transactions on]*, 47(9):2538–2541, 1999.

45. Kruse, F. A. and J.H. Huntington. "The 1995 Geology AVIRIS Group Shoot". *Summaries of the Sixth Annual JPL Airborne Earth Science Workshop*. 1996.

46. Landgrebe, David A. *Signal Theory Methods in Multispectral Remote Sensing*. John Wiley and Sons, 1 edition, 2003.

47. Lee, R. C. and W. E. Wright. "Development of Human Exposure-factor Distributions Using Maximum-entropy Inference". *Journal of Exposure Analysis and Environmental Epidemiology*, 4:329–341, 1994.

48. Lennon, M., P. Maupin, G. Grgoire, S. Prasherand J-A. Landry, and P. Goel. "Independent Component Analysis of Airborne Hyperspectral Data for the Study of Weed and Nitrogen Stress in Corn Crops". *Third European Conference on Precision Agriculture (ECPA)*. 2001.

49. Manolakis, D., D. Marden, and G. A. Shaw. *Hyperspectral Image Processing for Automatic Target Detection Applications*. Lincoln Laboratory Journal 1, Lincoln Laboratory, Lincoln Laboratory Massachusetts Institute of Technology Lexington, MA 01731, 2003.

50. Manolakis, D. and M. Rossacci. "Statistical Characterization of Natural Hyperspectral Backgrounds". *IGARSS Conference Proceedings*, 28. IGARSS, Denver, CO, 2006.

51. Manolakis, D. and G. Shaw. "Detection algorithms for hyperspectral imaging applications". *Signal Processing Magazine, IEEE*, 19(1):29–43, January 2002.

52. Manolakis, D. G. *Statistical Characterization of Natural Hyperspectral Backgrounds for Target Detection Applications*. Nasic remote sensing lecture series, MIT Lincoln Laboratories, Lexington, MA, May 2006.

53. Manolakis, D. G., G. A. Shaw, and N. Keshava. "Comparative Analysis of Hyperspectral Adaptive Matched Filter Detectors". *Proceedings, SPIE*, volume 4049, 2–17. SPIE, 2000.

54. Marden, D. and D. Manolakis. *Statistical Modeling of Hyperspectral Imaging Data and Their Applications*. Technical Report HTAP-16, Lincoln Laboratory, Massachusetts Institute of Technology, Lexington, Massachusetts, March 2004.

55. Marden, D.B. and D. Manolakis. "Modeling Hyperspectral Imaging Data". *Proceedings, SPIE 5093*, 253–262, 2003.

56. Marden, D.B. and D. Manolakis. ""Using Elliptically Contoured Distributions to Model Hyperspectral Imaging Data and Generate Statistically Similar Synthetic Data". *Proceedings, SPIE 5424*, 558–572. 2004.

57. Mardia, K. and P. Jupp. *Directional Statistics*. John Wiley & Sons, Ltd., 2000.

58. Mclachlan, G. and D. Peel. *Finite Mixture Models*. Wiley Series in Probability and Statistics. John Wiley & Sons, Inc., New York, 2000.

59. Meidunas, E. C., S. C. Gustafson, and D. G. Manolakis. "Robust Estimation of Tail-index Parameters with Application to Hyperspectral Data". Submitted to IEEE Transaction on Geoscience and Remote Sensing, 2006.

60. Montgomery, D.C. *Design and Analysis of Experiments, 5th Ed.* John Wiley & Sons, Inc, New York, 2001.

61. Neher, R. and A. Srivastava. "A Bayesian MRF Framework for Labeling Terrian Using Hyperspectral Imaging". *IEEE Transactions on Geoscience and Remote Sensing*, 43(6):1363–1374, June 2005.

62. Pickands, III, J. "Statistical Inference Using Extreme Order Statistics". *Annals of Statistics*, 3:119–131, 1975.

63. Randall B. Smith, Ph.D. *Introduction to Hyperspectral Imaging.* MicroImages, Inc, 11th Floor - Sharp Tower 206 South 13th Street Lincoln, Nebraska 68508-2010 USA, July 2006.

64. Rangaswamy, M., D. Wiener, and A. Ozturk. "Computer Generation of Correlated Non-gaussian Radar Clutter". *IEEE Transaction on Aerospace and Electronic Systems*, 31(1):106–116, January 1995.

65. Reed, I.S. and X. Yu. "Adaptive multiple-band CFAR detection of an optical pattern with". *Acoustics, Speech, and Signal Processing [see also IEEE Transactions on Signal Processing], IEEE Transactions on*, 38(10):1760–1770, 1990.

66. Reiss, R. D. and M. Thomas. "A New Class of Bayesian Estimators in Paretian Excess-of-Loss Reinsurance". *Astin Bulletin*, 29(2):339–349, 1999.

67. Rencher, A. *Methods of Multivariate Analysis, 2nd Ed.* Wiley Series in Probability and Statistics. John Wiley & Sons, New York, 2002.

68. Rousseeuw, P. J. and K. Van Driessen. "A Fast Algorithm for the Minimum Covariance Determinant Estimator". *Technometrics*, 41:212–223, 1999.

69. Rousseeuw, P.J. "Multivariate Estimation with High Breakdown Point". *Mathematical Statistics and Applications*, B:283–297, 1985.

70. Scharf, L. L. *Statistical Signal Processing Detection, Estimation, and Time Series Analysis,*. Addison-Wesley, 1991.

71. Scharf, L.L. and B. Friedlander. "Matched subspace detectors". *Signal Processing, IEEE Transactions on [see also Acoustics, Speech, and Signal Processing, IEEE Transactions on]*, 42(8):2146–2157, 1994.

72. Schirmacher, D., E. Schirmacher, and N. Thandi. "Stochastic Excess-of-Loss Pricing within a Financial Framework". *2005 Casualty Actuatial Society Forum.* 2005.

73. Schowengerdt, Robert A. *Remote Sensing, Models and Methods for Image Processing, 2nd Ed.* Academic Press, 1997.

74. Shoham, S., M. R. Fellows, and R. A. Normann. "Robust, Automatic Spike Sorting Using Mixtures of Multivariate t-distributions". *Journal of Neuroscience Methods*, 127:111–122, 2003.

75. Slifker, J. F. and S. S. Shapiro. "The Johnson System: Selection and Parameter Estimation". *Technometrics*, 22(2):239–246, 1980.

76. Solutions, ITT Visual Information. "ENVI - Environment for Visualizing Images". ITT Industries Inc., 2006. 4990 Pearl East Circle Boulder, CO 80301.

77. S.Tadjudin and D.A. Landgrebe. "Robust Parameter Estimation for Mixture Model". *IEEE Transactions on Geoscience and Remote Sensing*, 38(1):439–445, 2000.

78. Stein, D. W. J., S. G. Beaven, L. E. Hoff, E. M. Winter, A. P. Schaum, and A. D. Stocker. "Anomaly Detection from Hyperspectral Imagery". *Signal Processing Magazine, IEEE*, 19(1):58–69, January 2002.

79. Stuart, A. and J.K. Ord. *Kendalls Advanced theory of Statistics, Vol 1., Distribution Theory, Sixth Ed.* Kendalls Library of Statistics. John Wiley & Sons, Inc., 1994.

80. Thabane L., Drekic S. "Hypothesis testing for the generalized multivariate modified Bessel model". *Journal of Multivariate Analysis*, 86(2):360–374, August 2003.

81. Torbjrn Eltoft, Te-Won Lee, Taesu Kim. "On the Multivariate Laplace Distribution". *IEEE Signal Processing Letters*, 13(5):300–303, 2006.

1. REPORT DATE *(DD–MM–YYYY)*	2. REPORT TYPE	3. DATES COVERED *(From — To)*
21–12–2006	PhD Dissertation	Sept 2003 — Dec 2006

4. TITLE AND SUBTITLE

Robust Estimation of Mahalanobis Distances in Hyperspectral Images

5a. CONTRACT NUMBER

5b. GRANT NUMBER

5c. PROGRAM ELEMENT NUMBER

6. AUTHOR(S)

Eduardo C. Meidunas, Maj, USAF

5d. PROJECT NUMBER

5e. TASK NUMBER

5f. WORK UNIT NUMBER

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)

Air Force Institute of Technology
Graduate School of Engineering and Management
Bldg 640
2950 Hobson Way
WPAFB OH 45433-7765

8. PERFORMING ORGANIZATION REPORT NUMBER

AFIT/DS/ENG/07-02

9. SPONSORING / MONITORING AGENCY NAME(S) AND ADDRESS(ES)

Air Force Research Laboratory
Space Vehicles Directorate
Battlespace Environment Division
Dr Ron Lockwood
Hanscom Air Force Base, MA 01731
ronald.lockwood@hanscom.af.mil
(781) 981-0524

10. SPONSOR/MONITOR'S ACRONYM(S)

11. SPONSOR/MONITOR'S REPORT NUMBER(S)

12. DISTRIBUTION / AVAILABILITY STATEMENT

Approval for public release; distribution is unlimited.

13. SUPPLEMENTARY NOTES

14. ABSTRACT

This dissertation develops new estimation methods that fit Johnson distributions and generalized Pareto distributions to hyperspectral Mahalanobis distances. The Johnson distribution fit is optimized using a new method which monitors the second derivative behavior of exceedance probability to mitigate potential outlier effects. This univariate distribution is then used to derive an elliptically contoured multivariate density model for the pixel data. The generalized Pareto distribution models are optimized by a new two-pass method that estimates the tail-index parameter. This method minimizes the mean squared fitting error by correcting parameter values using data distance information from an initial pass. A unique method for estimating the posterior density of the tail-index parameter for generalized Pareto models is also developed. Both the Johnson and Pareto distribution models are shown to reduce fitting error and to increase computational efficiency compared to previous models.

15. SUBJECT TERMS

hyperspectral, stochastic image processing, Heavy-tailed Distributions,Extreme Value Theory

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
a. REPORT	b. ABSTRACT	c. THIS PAGE			Dr Steven C. Gustafson
U	U	U	UU	280	19b. TELEPHONE NUMBER *(include area code)* (937) 255–3636, ext 7227